

|  |
| --- |
| Connecticut Real Estate |
| 2011 - 2021 |
| December 2022  Aaron ‘Asa’ Sherwyn |

# Summary

## An exploration general trends within and the COVID-19 impact on Connecticut real estate sales.

The Connecticut Office of Policy and Management records all real estate sales greater than $2,000. Each entry records the date of sale, property address, town name, property type, sale price, and property assessments. This data can help forecast the value of properties across the state as well as assess the impact of the COVID-19 pandemic on Connecticut’s real estate market. The findings will be valuable to realtors as well as public policy makers.

|  |
| --- |
| Dataset This data is recorded in compliance with Connecticut General Statutes, section 10-26a and 10-26b. Information is released in grand lists years from October 1st through September 30th annually.  Publisher: data.ct.gov  Obtained from: <https://data.ct.gov/d/5mzw-sjtu>  License: Public domain  Last updated: October 20, 2022  Obtained on: December, 19 2022 |

# Data Profile

## Cleaning Process

| Step | Action | Description |
| --- | --- | --- |
| **Original dataset** |  | 997,213 rows and 14 columns. |
| Relevance checks | Dropped unnecessary columns and rows | Dropped Non Use Code, Assessor Remarks, OPM remarks, and Location columns as well as List Years 2001-2009. Dropped 488,808 rows and 4 columns. Current dataset: 508.405 rows and 10 columns. |
| Type checks | Changing columns to correct data types. | Date Records changed to datetime type with yyyy-mm-dd format. Address, Property Type, and Residential Type had mixed data types; converted to strings. |
| Consistency checks | Renaming columns | Old Name' : 'new\_name' 'Serial Number' : 'serial\_number' 'List Year' : 'list\_year' 'Date Recorded' : 'date\_recorded' 'Town' : 'town' 'Address' : 'address' 'Assessed Value' : 'assessed\_value' 'Sale Amount' : 'sale\_amount' 'Sales Ratio' : 'sales\_ratio' 'Property Type' : 'property\_type' 'Residential Type' : 'residential\_type' |
| Consistency checks | Mixed structures/ formats | All structures/formats seem consistent and accurate. |
| Outlier checks | Removed outlier values from the dataset. | Outliers seem to have unusually large or small sales\_ratios.   "The Balance" explains that assessed value is almost always lower than market price. Thus, a sales\_ratio should be between 0 and 1, usually staying above 0.5. Due to the unusual sales during the COVID pandemic, some sales\_ratios above 1 should be allowed. (https://www.thebalancemoney.com/assessed-value-vs-market-value-what-s-the-difference-5197369#:~:text=The%20tax%20assessor%20for%20your,the%20assessed%20value%20is%20%24225%2C000)  We can structure a preliminary outlier formula of a sales\_ratio being greater than Q3+2IQR or less than Q1-2IQR (i.e., sales\_ratio > 1.40845 or sales\_ratio < 0.0467).  sales\_amount and assessed\_value now adhere more closely to a trend line as expected. A few outliers still exist in records with a sales\_amount > 300,000,000.  Removing records with sales\_amounts > 300,000,000 normalized the set nicely.  Dropped 45,947 rows Current dataset: 462,458 rows and 10 columns |
| Missing values checks | Looking for entries with NaN values | address had 3 NaN values - should not interfere with analysis, so left as is. property\_type had 27279 NaN values all of which also had NaN residential\_types. All property types with NaN values were dropped due to insufficient categorical information. residential\_type had 31965 NaN values. 27279 were dropped due to lack of property\_type. Remaining 4,686 are non-residential types, thus should not have residential\_type value and were left as is.  Dropped 27,279 rows. Current dataset: 435,179 rows and 10 columns |
| Duplicates checks | Checking for duplicated entries. | No duplicate entries found. |
| Contradictions checks | Checking for contradictor data. | residential\_type is a subcategory of property\_type: Residential. residential\_has Single Family, Condo, Two Family, Three Family, and Four Family; property\_type also has Single Family, Condo, Two Family, Three Family, and Four Family. property\_type: Condo has only one residential\_type which is Condo, property\_type: Single Family has only one residential\_type which is Single Family, etc. This is redundant and unhelpful in the analysis.  property\_types of Single Family, Condo, Two Family, Three Family, and Four Family changed to "Residential". |
| **Final dataset** |  | 435,179 rows and 10 columns |

## Column Variables

|  |  |  |  |
| --- | --- | --- | --- |
| Column Variable | Description | Datatype | Time Variant |
| serial\_number | Unique record ID | Qualitative, nominal | No |
| list\_year | Year classification for the record | Quantitative,  discrete | No |
| date\_recorded | Exact date the real estate sold | Quantitative, continuous | No |
| town | Name of the town the real estate is in | Qualitative, nominal | No |
| address | Street address of the real estate | Qualitative,  nominal | No |
| assessed\_value | Tax-assessed value (in USD) on record with local municipality | Quantitative,  continuous | No |
| sale\_amount | Exact amount (in USD) the real estate sold for | Quantitative,  continuous | No |
| sales\_ratio | assessed\_value divided by sale\_amount | Quantitative,  continuous | No |
| property\_type | Category of the real estate | Qualitative,  nominal | No |
| residential\_type | Sub-category of Residential property\_types | Qualitative, nominal | No |

## Limitations

The data this set provides can reveal general trends about real estate sales in Connecticut. It does not, however, provide detailed enough data to determine specific trends. For this it would need square footage, number of rooms/bathrooms, zip code, and days listed for each property. Sales should be normalized by the state’s annual population and town populations if possible.

## Ethics

As the data comes from a government source and via obligatory reporting, it should be reliable. The cleaning process did show that reporting was not always consistent across the state which caused some discrepancies. There is enough personally identifiable information in this data set to determine who sold the property and to whom. The analysis should be conducted with care and with the consideration of those it might implicate. The exact address of each record could be removed to help protect those involved, but as all this data is already public record it would hardly deter a motivated individual.

# Exploratory Questions

This analysis is dual purposed: first, to determine and forecast the general sales trends within the State of Connecticut and its individual towns. And second, to explore COVID-19 impact on the real estate market in Connecticut.

## General Sales Trends:

1. How have sales changed from year to year across all property types?
2. How have sales changed from year to year for each individual property type?
3. Do sales vary from month to month?
   1. When do most sales occur? And the least?
   2. What month has the lowest sales ratio? And the highest?
   3. Is there a seasonality to real estate sales?
4. Do sales volume (count of sales), sales amounts, and sales ratios vary by town?
   1. What are the top 5 towns for volume of sales? And the bottom 5?
   2. What are the top 5 towns for total amounts of sales? And the bottom 5?
   3. What are the top 5 towns for average sales ratio? And the bottom 5?
   4. In which towns have sales ratios been increasing? And decreasing?
5. Do sales volume (count of sales), sales amounts, and sales ratios vary by property type or residential type?
   1. What type has the largest volume of sales? And the smallest?
   2. What type has the largest total amount of sales? And the smallest?
   3. What type has the smallest average sales ratio? And the largest?
   4. In which types have sales ratios been increasing? And decreasing?
6. Forecast trends for the state and each property type.

## COVID-19 Impact:

Follow the same structure as General Sales Trends, but break each question into pre- and peri-pandemic chronological categories and compare the two. March 11, 2020 will be the dividing line (<https://www.who.int/news/item/27-04-2020-who-timeline---covid-19>).