Real Estates in Connecticut

Data Curation by Anonymous

# General Dataset Information:

File Name: Real\_Estate\_Sales\_2001-2020\_GL.csv

Description: Property Sales, Assessments, and Trends in Connecticut 2001-2020

**Dataset Details**: 997213 Rows & 14 Columns

**Size**: 107,807KB ( 105MB )

**Source**: Kaggle - [Dataset Link](https://www.kaggle.com/datasets/utkarshx27/real-estate-sales-2001-2020-gl)

# Data Profile:

* Apply Excel filter to the dataset.
* Examine each attribute's unique value for inconsistency.
* Notable Features
  + 169 Unique Towns Names
  + 15 Unique Years
  + Property Type - Residential, Single Family, Apartments, Condo, Four Family, Three Family, Two Family
  + Residential Type - Single Family, Condo, Two Family, Three Family, Four Family
  + Assessor Remarks - Contain property sales reasoning such as Family Sales, Foreclosure, Will, Zoning, Plottage, Personal Property Exchange, etc.

**Recorded Inconsistency:**

* Address holding numerical, date, and blank values.
* Property Type holding blank values
* Residential Type holding blank values

# 

# Data Wrangling:

* Sorted Address attribute to isolate the blank, numerical and date value. The record is removed and inserted into the “Removed” worksheet. Total 52 records were removed.
* Feature Engineered “Property Type Adjusted” column, substitute any blank value with “Not Listed”
* Feature Engineered “Residential Type Adjusted” column, substitute any blank value with “Not Listed”
* Feature Engineered “Year Recorded” column, extracted the year from the “Date Recorded” attribute.
* Feature Engineered “Sales Bucket” column, segregate “Sales Amount” attribute into generic price range bins. There are 5 bins:
  + 0 - 100000
  + 100k - 200k
  + 200k - 500k
  + 500k - 1M
  + 1M+
* Feature Engineered “Correct Bins” column, segregate “Sales Amount” attribute into specific price range bins. There are 9 bins:
  + 0 - 100000
  + 100k - 200k
  + 200k - 300k
  + 300k - 400k
  + 400k - 500k
  + 500k - 600k
  + 600k - 700k
  + 700k - 800k
  + 1M+
* Filter “Property Type” and “Residential Type” for any records that contain specific Residential Features - Residential, Single Family, Apartments, Condo, Four Family, Three Family, Two Family.
* Isolate Residential records and place it in a new worksheet “**Residential Datasets**”. 609,328 records were isolated.

**Data Table Schema:**

| Field | Type | Description |
| --- | --- | --- |
| Serial Number | NUMBER | A unique identifier for each record in the dataset. |
| List\_Year | NUMBER | The grand list year in which the sale was recorded. |
| Date\_Recorded | DATE | The date when the sale was recorded. |
| Town | STRING | The town where the property is located. |
| Address | STRING | The address of the property. |
| Assessed\_Value | NUMBER | The assessed value of the property. |
| Sale\_Amount | NUMBER | The sales price of the property. |
| Sales\_Ratio | FLOAT | The sales ratio of the property. |
| Property\_Type | STRING | The type of the property (residential, apartment, commercial, industrial, or vacant land). |
| Residential\_Type | STRING | The type of residential property (if applicable). |
| Non\_Use\_Code | STRING | The non-use code associated with the property (if applicable). |
| Assessor\_Remarks | STRING | Remarks or comments provided by the assessor (if available). |
| OPM\_Remarks | STRING | Remarks or comments provided by the Office of Policy and Management (if available). |
| Location | STRING | The location of the property (if available). |