# Data source

## The Data Source name:

1. **Real Estate Sales 2001-2020 GL**. State of Connecticut (<https://catalog.data.gov/dataset/real-estate-sales-2001-2018> )

## Compliance with the requirements:

The data set is:

Open source;

Comes from an authentic/authoritative source;

Includes non-anonymized column names;

is no more than three years old (up to a maximum of 10 years if you’ve found a perfect data set

for your needs, and no newer data is available);

Contains at least two continuous variables (excluding index or ID variables, dates, years, etc.);

Contains at least two categorical variables (excluding index or ID variables, dates, years, etc.);

Contains at least 1,500 rows;

Includes a geographical component with at least two different values (e.g., countries, continents,

U.S. states, cities, latitude, and longitude values—anything is to visualize on a map)

## The summary

**Real Estate Sales 2001-2020 GL, updated** [**August 12, 2023**](https://catalog.data.gov/dataset/real-estate-sales-2001-2018#sec-dates)

The Office of Policy and Management maintains a listing of all real estate sales with a sales price of $2,000 or greater that occur ***between October 1 and September 30*** of each year. For each sale record, the file includes town, property address, date of sale, property type (residential, apartment, commercial, industrial, or vacant land), sales price, and property assessment.

Data are collected following Connecticut General Statutes.

Annual real estate sales are reported by grand list year (October 1 through September 30 each year). For instance, sales from 2018 GL are from 10/01/2018 through 9/30/2019.

## The data cleaning

The cleaning included:

* Dropping non-relevant columns
* Checking the Hypothesis: if 'Property Type' ='Commercial', then 'Residental Type' =NaN.

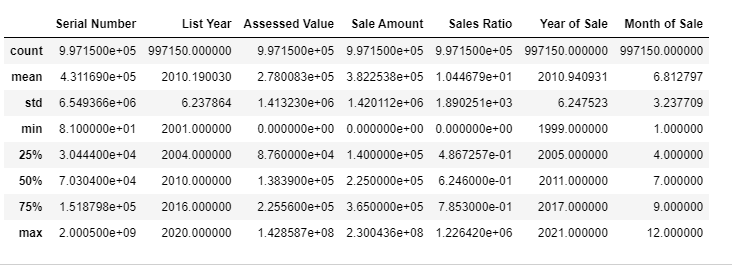
Therefore, creating a subset for commercial Property Type and counting NaNs in columns

* Substitution of NaNs in 'Property Type' and 'Residential Type' with 'Non-Indicated'
* Fixing one outlier
* Creating columns Year and Month instead of List Date
* Checking data types

## The data understanding

1. **The data for completeness, integrity, and consistency.**
2. The data is complete: all listed objects are there even if they were not sold.
3. The data is integer: all collected data is presented
4. The original data is not consistent: the NaNs emerged because of some columns by sold RE relevance only and can be substituted with the string “non-indicated.”

After cleaning the statistics of the dataset as follows:



The STD value by the ‘Assessed Value’ and the ‘Sales Amount’ columns is high because of the assignation of 0 values if the real estate wasn’t sold.

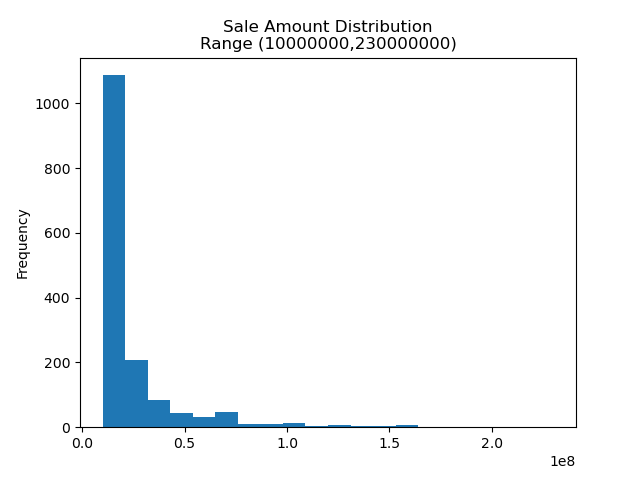
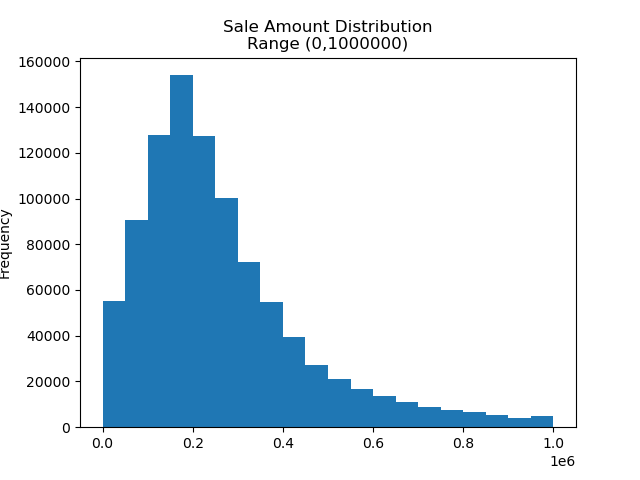
The max value of the serial number column is the outlier, but it was not corrected because we didn't have the goal to identify the property.

1. **The distribution of Sales Amount data**

The distribution of Sales Amount data by sod RE is high left-skewed. That means most sold RE objects have lower values.

Figure1. Sale Amount distribution, Figure 2. Sale Amount distribution,

Range: 0 - 1mln $ Range: 1mln -20 mln$



The RE objects were registered starting from 2000$. Figure 1 shows the max frequency at 200,000$. The Sales Amount of 225,000$ corresponds to the Median of the dataset. Figure 2: confirm the left-skewed distribution showing the tail of the shape.

The calculated P-value is 0.0 by α=0.05; this result confirms the distribution is not normal.

1. **The z-test and outliers by sold RE**

Conducted z-test by outlier\_threshold = 3 standard deviation identified **4390** outliers from **995367** Sale Amount values considerating Sale amount >0. The quantity of identified outliers is **0.44%**, less than 5%, and can be accepted. The outliers were not deleted from the database because they didn’t change the shape of the distribution.

1. **The specific tests for non-normal distribution were also conducted:**

Sign Test:

Test Result: BinomTestResult(k=491665, n=987947, alternative='two-sided', statistic=0.49766333619111147, pvalue=3.4159452041092726e-06)

Wilcoxon Signed-Rank Test:

Test Statistic: 205836116759.5

P-Value: 0.0

Both tests give a p-value less than α=0.05;

## Considerating limitations and ethics

Ethical constraints may arise if there is a potential for identifying property owners based on published data. Since the owners' names have been removed, the possibility of identification remains only in the case of unique real estate properties. The only property with an exceptionally high value has been removed from the database.

# The question Definition

1. Identify the characteristics of real estate that contribute to its profitability when sold.

2. Identify the type of real estate that is more readily accessible for sale.

3. Identify areas with a uniform level of property pricing.

4. Determine if there are variations in commercial and residential real estate sales, and if so, what are they?

5. Determine the optimal investment strategy for real estate in the region for various investment amounts.