IE6600: Data Analysis and Visualization Project Report

Project 1: Connecticut Real Estate Analysis

Madison Rodriguez, Adam Luk, Ameya Patil, Leela Maunika Talluri

1. Introduction

1.1. Objective

The primary objective of this project is to analyze the trends in real estate prices in Connecticut from 2001 to 2021 and examine possible relationships with local unemployment statistics during the same period. The analysis aims to reveal patterns and correlations between real estate dynamics in relation to economic factors like unemployment.

1.2. Dataset Description

The chosen dataset from data.gov was Real Estate Sales 2001-2021. This dataset includes real estate data for Connecticut including information about sale prices, sale dates, property types and locations. The additional dataset from data.gov is Local Area Unemployment Statistics. This provides unemployment statistics such as unemployment rates, labor force participation and employment population for approximately 7,000 areas, including Connecticut.

2. Data Acquisition and Preparation

2.1. Initial Inspection

As part of the initial inspection the team evaluated the first dataset which we defined as df ('RealEstate_Sales.csv') using shape, describe(), and info(). The dataset contained 1,054,159 data points and had 14 columns. It consisted of the data type int64, object and float64. The first 8 columns did NOT have any null values, while the other 6 did.

The second dataset was defined as df2 ('unemp_stats.csv') and was evaluated using the same functions. The dataset contained 150,336 data points and had 11 columns. It consisted of the data type int64, object and float64 as well. This dataset was complete and had NO null values.

2.2. Data Cleaning

The first and second steps for the Real Estate dataset involved addressing missing data and removing duplicates. This was accomplished using the dropna() and drop_duplicates() functions to eliminate any rows with all null values or repeated entries. Since the shape of the dataset remained unchanged before and after this process, no entirely null or duplicate rows were present. However, some rows had partial missing values, which were filled using the mode() function for each column, as all columns with missing values were categorical. Additionally, the ‘Location’ column was dropped since the team was only interested in the ‘Town’ column, which was available for every data point, rather than the exact location.

The third step was to convert and normalize the data. The only conversion required in both datasets was the date column, which was handled using the to_datetime() function. The team decided that encoding categorical data was unnecessary. However, after cleaning the data, the team used the info() function to confirm that the data types were correctly changed and that there were no null values present in the dataset, as shown in Figure 1.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1054159 entries, 0 to 1054158
Data columns (total 14 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   Serial Number     1054159 non-null  int64
 1   List Year         1054159 non-null  int64
 2   Date Recorded     1054157 non-null  object
 3   Town              1054159 non-null  object
 4   Address           1054108 non-null  object
 5   Assessed Value    1054159 non-null  float64
 6   Sale Amount       1054159 non-null  float64
 7   Sales Ratio       1054159 non-null  float64
 8   Property Type     671713 non-null   object
 9   Residential Type  660275 non-null   object
 10  Non Use Code      302242 non-null   object
 11  Assessor Remarks  161472 non-null   object
 12  OPM remarks       11564 non-null    object
 13  Location          254643 non-null   object
dtypes: float64(3), int64(2), object(9)
memory usage: 112.6+ MB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1054159 entries, 0 to 1054158
Data columns (total 13 columns):
 #   Column            Non-Null Count    Dtype
---  ------            --------------    -----
 0   Serial Number     1054159 non-null  int64
 1   List Year         1054159 non-null  int64
 2   Date Recorded     1054157 non-null  datetime64[ns]
 3   Town              1054159 non-null  object
 4   Address           1054108 non-null  object
 5   Assessed Value    1054159 non-null  float64
 6   Sale Amount       1054159 non-null  float64
 7   Sales Ratio       1054159 non-null  float64
 8   Property Type     1054159 non-null  object
 9   Residential Type  1054159 non-null  object
 10  Non Use Code      1054159 non-null  object
 11  Assessor Remarks  1054159 non-null  object
 12  OPM remarks       1054159 non-null  object
dtypes: datetime64[ns](1), float64(3), int64(2), object(7)
memory usage: 104.6+ MB
```

Figure 1: df.info() Before and After Data Cleaning

The Unemployment dataset followed the same steps 1-3, however did not have any NA values so no values needed to be filled. It also had one date column which was converted using the to_datetime() function.

```
Data columns (total 11 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   State FIPS code      150336 non-null  int64
 1   State abbreviation   150336 non-null  object
 2   State name           150336 non-null  object
 3   Series ID            150336 non-null  object
 4   Metric code          150336 non-null  int64
 5   Year                 150336 non-null  int64
 6   Period code          150336 non-null  object
 7   Period name          150336 non-null  object
 8   Metric name          150336 non-null  object
 9   Value                150328 non-null  float64
 10  Update date          150336 non-null  object
dtypes: float64(1), int64(3), object(7)
memory usage: 12.6+ MB
```

```
Data columns (total 11 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   State FIPS code      150336 non-null  int64
 1   State abbreviation   150336 non-null  object
 2   State name           150336 non-null  object
 3   Series ID            150336 non-null  object
 4   Metric code          150336 non-null  int64
 5   Year                 150336 non-null  int64
 6   Period code          150336 non-null  object
 7   Period name          150336 non-null  object
 8   Metric name          150336 non-null  object
 9   Value                150328 non-null  float64
 10  Update date          150336 non-null  datetime64[ns]
dtypes: datetime64[ns](1), float64(1), int64(3), object(6)
memory usage: 12.6+ MB
```

Figure 2: df2.info() Before and After Data Cleaning

3. Exploratory Data Analysis (EDA)

3.1. Overview

Through our analysis, we aimed to determine if there was any correlation between the unemployment rate in Connecticut and real estate sales. Specifically, we focused on towns with the highest and lowest total sales and mean sale value during this period. Additionally, we examined the sales trends of different residential types over the years and compared the assessed values to the sale prices for each residential type. Lastly, we explored the correlation between assessed and realized sale amounts.

3.2. Visualizations

3.2.1 Bar Plots

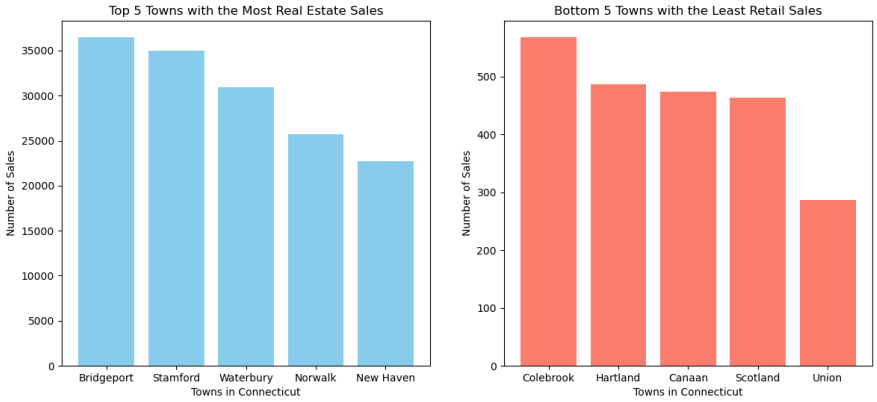

Figure 1: These bar graphs depict the towns with the most and least real estate sales between 2001 and 2021. Notably, Bridgeport, Stamford, Waterbury, Norwalk, and New Haven emerged as the leading towns, while Colebrook, Hardland, Canaan, Scotland, and Union recorded the lowest amount of sales. The towns with the highest sales ranged approximately from 24,000 to 35,000, while the lowest sales ranged from 300 to 550 real estate sales.

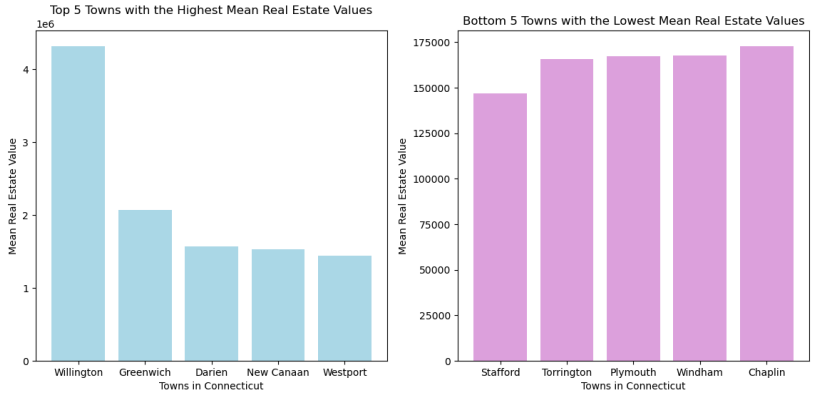

Figure 2: These bar graphs depict the average real estate values per town, highlighting the highest and lowest mean values in Connecticut based on data from 2001 to 2022. Notably, leading towns like Willington, Greenwich, Darien, New Canaan, and Westport demonstrate high average values, whereas Stafford,

Torrington, Plymouth, Windham, and Chaplin show comparatively lower averages. The top-ranking towns have values ranging from approximately $1.4 million to $4.3 million, while the lowest-ranking towns average between $146,000 and $175,000 in real estate prices.

3.2.2 Line Graphs

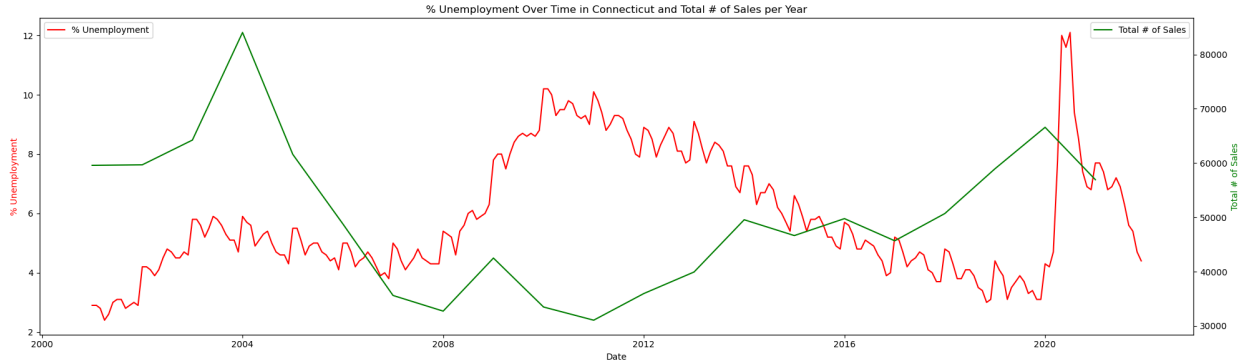

Figure 3: This line graph illustrates the percentage of unemployment and the total number of sales in Connecticut, both plotted on the same graph. Significant spikes in the total number of sales are observed in 2004 and 2020. Additionally, there are notable peaks in unemployment around the years 2010 and 2020.

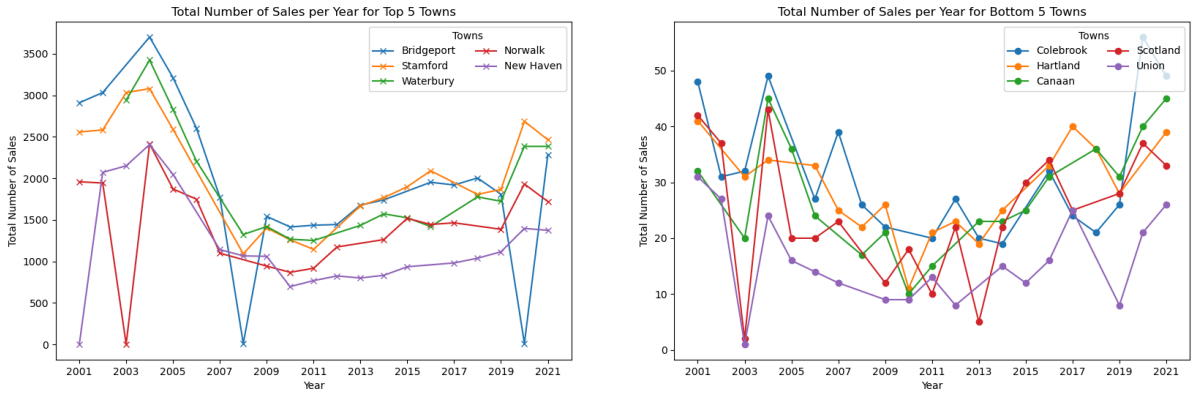

Figure 4: This figure illustrates the total sale amount over time for the top 5 and bottom 5 towns in Connecticut. It reveals notable trends in real estate activity between the years 2004 and 2020. There was a distinct spike in sales around 2004, followed by a substantial decline until 2011. After 2011, there is a gradual increase in sales activity, reaching a peak around 2020. This pattern is consistent across both the towns with high sale amounts and those with lower sale amounts, although the lower towns have more sporadic fluctuations

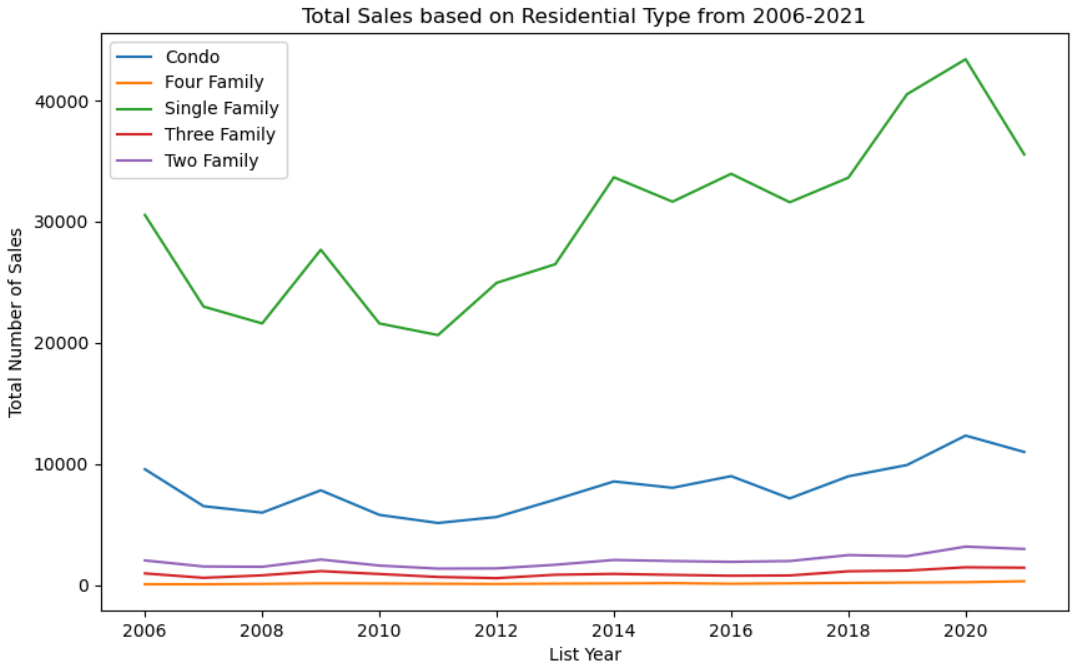

Figure 5: This line plot analyzes the changes in real estate sales of Residential Types in Connecticut from 2006 to 2021. There's a small peak in 2009, with a gradual increase observed until 2020. Notably, single-family homes and condos stand out due to their higher sales volumes during this period.

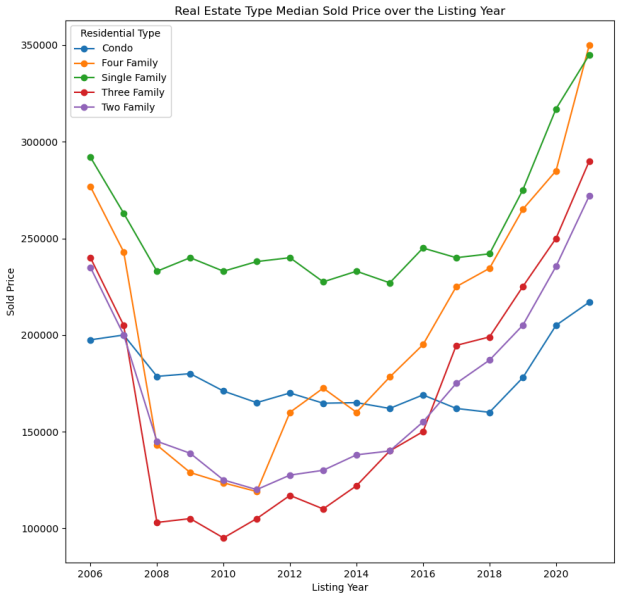

Figure 6: The line graph shows the median sold prices for various types of residential properties over the listing year. In 2021, the median sold prices for single-family and four-family homes rose to nearly $350,000, making them the highest. Meanwhile, three-family homes, two-family homes, and condos had the third, fourth, and fifth highest median sold prices, respectively. Overall, the median sold prices for all residential property types increased over the listing year.

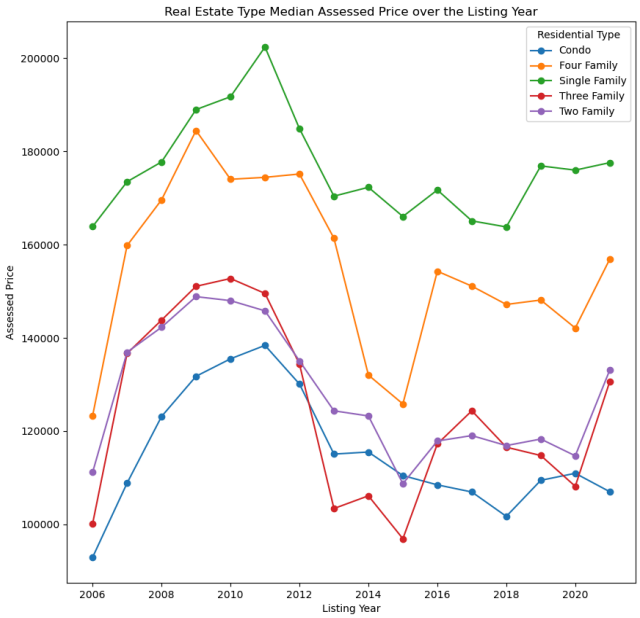

Figure 7: The line graph indicates that single-family homes consistently have the highest assessed prices, while condos have the lowest compared to other property types. Additionally, it appears that house prices for all residential types peaked between 2008 and 2010.

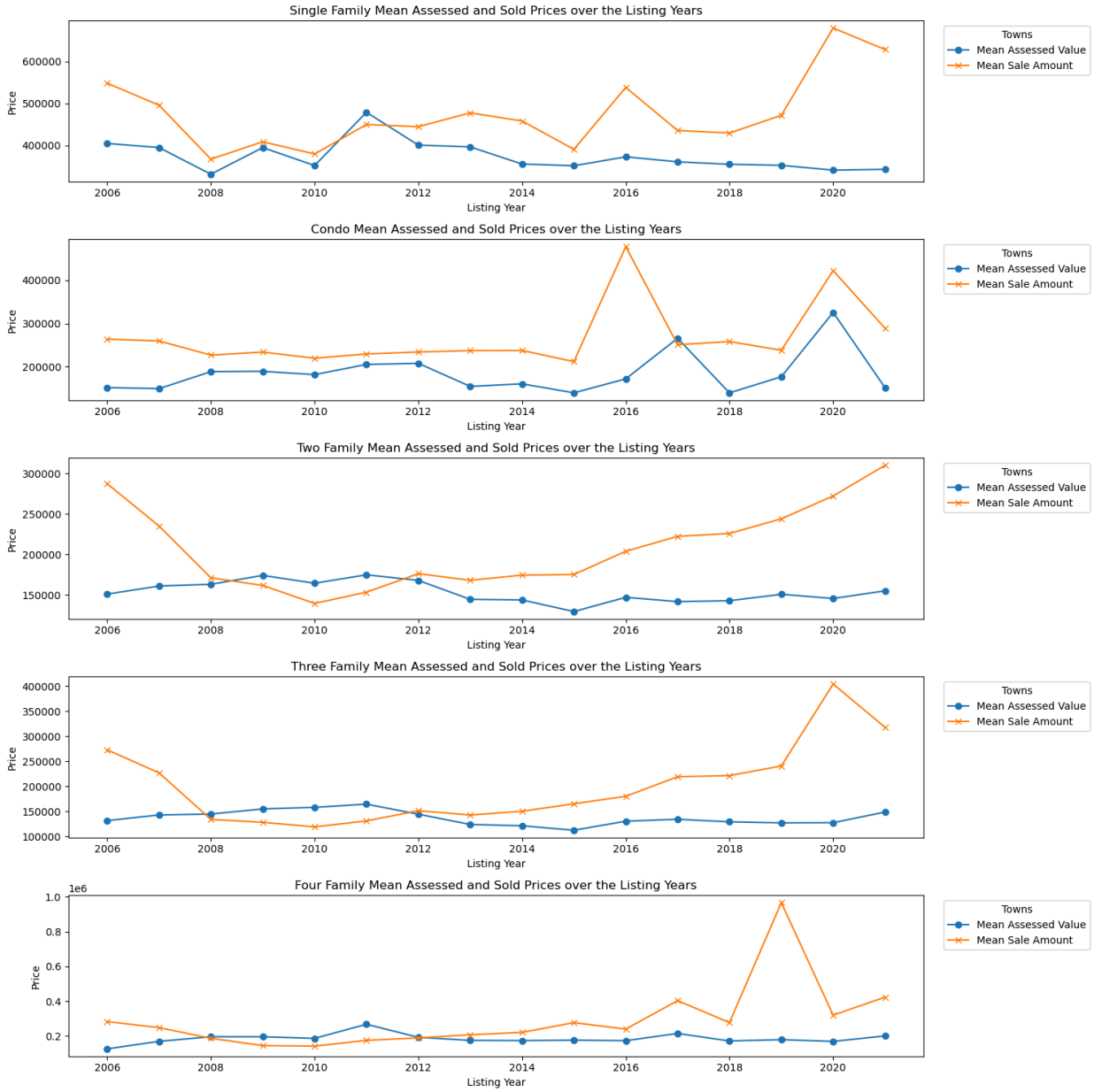

Figure 8: These line plots illustrate the comparison between assessed values and sale values based on 'Residential Type' in Connecticut from 2006 to 2021. Across all the plots, there is a noticeable peak in Mean Sale Amount in 2020, with the exception of Four Family homes. Additionally, between 2008 and 2012, it was the only period when the Mean Assessed Value surpassed the Mean Sale Amount for Two, Three, and Four Family homes. Single family and Condo Mean Sale Amounts also reached a peak in 2016. Throughout these Residential Types, there is a consistent upward trend from the low point in 2008 to 2020.

3.2.3 Pie Chart

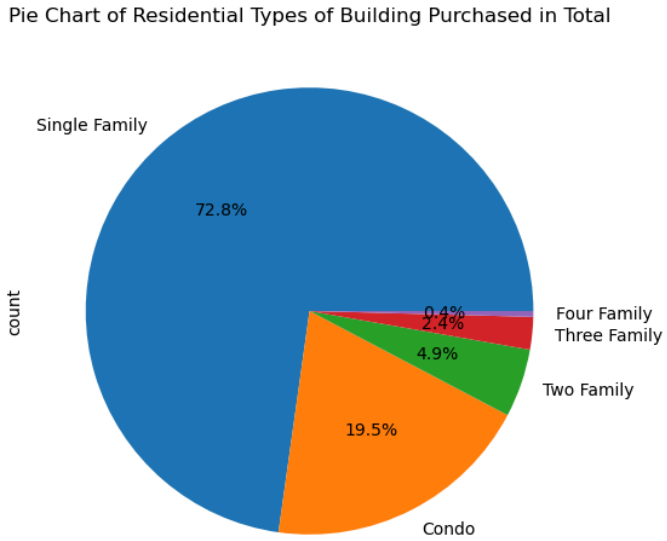

Figure 9: This chart shows the Residential Types of real estate purchased. The pie graph of all residential types purchased show that single family is the highest count as it comprised 72.8% of the properties sold. While condos were the second most purchased with 19.5%.

3.2.4 Scatter Plot

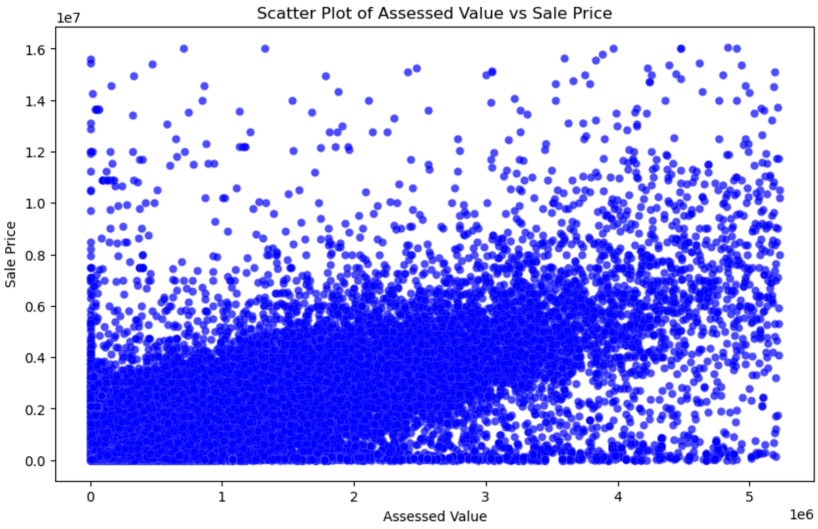

Figure 10: The scatter plot visualizes the relationship between 'Assessed Value' and 'Sale Amount', revealing a positive correlation where higher assessed values generally correspond to higher sale prices. While a slight upward trend is observed, a notable clustering of data points in the bottom-left corner indicates a segment of properties with lower assessed values and sale prices.

3.3. Statistical Analysis

Through describe() the following figure shows the count, mean, std, etc. From the statistics about the dataset, the mean assessed value is around $280,000 compared to the mean sale value of around $400,000. From figures 6, 7, 8, 10, and 11, it can be concluded that the sale amount is dramatically higher than the assessed values of Connecticut property.

```
Statistics about the dataset:
       Serial Number     List Year  Assessed Value   Sale Amount   Sales Ratio
count   1.054159e+06  1.054159e+06    1.054159e+06  1.054159e+06  1.054159e+06
mean    5.027140e+05  2.010774e+03    2.797416e+05  3.990286e+05  9.953241e+00
std     7.230239e+06  6.540711e+00    1.650117e+06  5.229758e+06  1.838434e+03
min     0.000000e+00  2.001000e+03    0.000000e+00  0.000000e+00  0.000000e+00
25%     3.055200e+04  2.004000e+03    8.845000e+04  1.422000e+05  4.816008e-01
50%     8.008000e+04  2.011000e+03    1.395800e+05  2.300000e+05  6.162887e-01
75%     1.608155e+05  2.017000e+03    2.270000e+05  3.700000e+05  7.764000e-01
max     2.000500e+09  2.021000e+03    8.815100e+08  5.000000e+09  1.226420e+06
<class 'pandas.core.frame.DataFrame'>
```

Figure 11: Statistical description on Assessed and Sale Value of Connecticut Sales

Additionally, the team did further analysis looking into the median values of Assessed Value and Sale Value. When specifically comparing the median of each year, it shows that the sale price can go up to 124% greater than the assessed. Interestingly it also illustrates the United States 2008 recession showing a decrease in sale price comparatively to the assessed from 2008 to 2012.

List Year	
2006	124.452555
2007	52.055566
2008	-15.688933
2009	-30.167493
2010	-29.031146
2011	-31.781701
2012	-8.652337
2013	6.893881
2014	21.193759
2015	41.665011
2016	26.377187
2017	48.908008
2018	59.339539
2019	78.908993
2020	100.513687
2021	123.100459

Figure 12: % Difference between Assessed and Sale Value

4. Conclusion

4.1. Summary of Findings

The two bar plots indicate that there is no direct correlation between the towns with the highest and lowest number of real estate sales and those with the highest and lowest mean real estate values. This challenges that higher sales volumes directly translate to higher property values. It suggests that factors beyond sale frequency play a significant role in determining real estate values in different towns. Moreover, there's notable skewing in the mean real estate value for Willington. After investigation it was mainly due to sales of apartment buildings exceeding $10

million. Further investigation is needed to determine if this anomaly exists across all towns or if it's specific to certain towns.

After comparing both datasets there seems to be a visible pattern suggesting an inverse relationship between the percentage of unemployment and the total volume of real estate sales. This suggests that periods of high unemployment rates often coincide with lower levels of real estate sales, while periods of low unemployment tend to see higher levels of sales activity. This relationship may be attributed to various factors such as job security, and overall economic stability. However, there is an exception to this trend in 2020, most likely due to the pandemic that caused both unemployment rates and real estate sales to be high. The data in Figure 4 further confirms an inverse relationship between the unemployment rate and real estate sales, as evidenced by the trends observed in both the top 5 and bottom total sales towns.

One significant constraint is the lack of Residential Type data from 2001 to 2006, which was filtered and excluded in the data to create Figures 5-9, which heavily relies on Residential Type grouping. In Figure 5, all the Residential Types showed a small decline between 2006 and 2012, with occasional small peaks, followed by a continuous increase up to 2020, mirroring trends observed when data is grouped by towns. Similarly, the Mean Sale Amount for most Residential Types exhibited a decrease from 2006 to 2010, followed by an increase up to 2020. However, there was a notable spike in data for Condo Sales in 2016. These observations suggest that there may be underlying trends in the real estate market, with fluctuations in both Residential Type data and Mean Sale Amounts over time. Additionally, the significant peak in Condo Sales data in 2016 may signify a unique event or trend specific to that housing category during that period.

Lastly, the scatter plot illustrates the relationship between Assessed and Sale Amount, highlighting how high unemployment rates may result in decreased demand and lower sale prices in the real estate market. Conversely, periods of economic growth and low unemployment rates tend to increase property values.

4.2. Implications

The findings from the analysis provide valuable insights into various aspects of the real estate market. Firstly, the lack of a direct correlation between towns with the highest sales volumes and those with the highest mean property values suggests that other factors play a significant role in determining real estate values across different areas. Also, the observed inverse relationship between unemployment rates and real estate sales underscores the influence of economic factors on market activity, with high unemployment rates typically associated with lower sales volumes and vice versa. Except for the anomaly observed in 2020, where both unemployment rates and

real estate sales were high. Additionally, the analysis of Residential Type data reveals underlying trends in the market, with fluctuations in both sales volumes and mean sale amounts over time. Lastly, the scatter plot depicting the relationship between Assessed and Sale Amounts highlights the impact of economic conditions on property demand and prices, emphasizing how economic downturns can lead to decreased demand and lower sale prices, while periods of economic growth tend to drive up property values.

4.2. Future Work

In addition to the current analysis, it would be beneficial to explore the impact of other economic indicators on real estate sales in Connecticut. Factors such as interest rates, population demographics, housing inventory levels, and local market conditions could provide valuable insights into the dynamics of the real estate market. Furthermore, conducting a comparative analysis with neighboring states or national trends could offer a broader perspective on Connecticut's real estate landscape. Additionally, considering qualitative factors such as changes in lifestyle preferences, urban development initiatives, and transportation infrastructure improvements could further enrich the analysis.