DATA ANALYSIS RESEARCH PROJECT

VISHAL ORSU

APRIL 8, 2023

AIT580 FINAL PROJECT
GEORGE MASON UNIVERSITY

Contents

I ABSTRACT 1

II INTRODUCTION 1

III LITERATURE REVIEW 2

IV METHODOLOGY 3

V RESULTS AND ANALYSIS 4

***A. Univariate Analysis**:* 4

B. Multivariate Analysis 6

C. Research Question Analysis 9

D. SQL Analysis (Queries and Output) 16

VI CONCLUSION, LIMITATIONS AND FUTURE RESEARCH: 19

REFERENCES 20

Table OF Figures

Fig. 1. Bar Plot of top 20 towns (Nominal Data) 4
Fig. 2. Bar Plot of Sale Rankings (Ordinal data) 5
Fig. 3. Histogram of Latitude and Longitude (Interval type) 5
Fig. 4. Boxplot of Sales Ratio (Ratio type) 6
Fig. 5. Correlation Plot (Heatmap) 6
Fig. 6. Correlation of three variables Assessed.Value, Sale.Amount, and Sales.Ratio 7
Fig. 7. Scatter Plot of Sale Amount and Assessed Value 7
Fig. 8. Regression Model 8
Fig. 9. Linear Regression Plot 9
Fig. 10. Real Estate Sale Price from 2001-2020 in Connecticut 10
Fig. 11. Line plot of Sales Volume from 2001-2020 10
Fig. 12. Real Estate Sales Volumes and Prices over 2001-2020 period in Connecticut 11
Fig. 13. Bar Plot of Most Common Property Types Sold in Connecticut from 2001-2020 12
Fig. 14. Line Plot of Average Sale Price of Residential Properties over time 13
Fig. 15. Average Sale Amount of Top 10 Residential Towns Connecticut 14
Fig. 16. All Real Estate Properties Locations in Connecticut from 2001-2020 15
Fig. 17. Geographic Distribution of Real Estate Sales from 2001-2020 in Connecticut 16
Fig. 18. SQL query displaying First 10 Rows of the CSV file 17
Fig. 19. SQL query displaying Sale Ranking Count 17
Fig. 20. SQL query displaying all rows with town filtered as Greenwich 18
Fig. 21. SQL query showing Average Sale amount through 2001 to 2020 18
Fig. 22. SQL Query displaying total count of Residential type properties 18

GLOSSARY

Real Estate	Property in the form of land, buildings, or other structures that are used or intended to be used for residential, commercial, or industrial purposes.
Sales Volume	The total number of real estate transactions in a particular period.
Sales Price	The amount of money paid for a property in a real estate transaction.
Property Type	The classification of a property based on its intended use, such as residential, commercial, or industrial.
Condominium	A type of residential property where each unit is individually owned, but the common areas are owned collectively.
Commercial Property	Property that is used for business or commercial purposes.
CSV	Comma Separated Values
Economic Downturn	A period of economic decline or recession, typically characterized by a decrease in sales volumes and prices in the real estate market.
Zoning Regulations	Laws and regulations that govern the use of land and structures in a given area, typically established by local government.
Policy Interventions	Actions taken by government or other organizations to influence or regulate market behavior, such as tax incentives or zoning regulations.

Data Analysis of Real Estate Sales from 2001 to 2020 in Connecticut

Vishal Orsu · AIT580 Final Project

I ABSTRACT

Real estate is a crucial and dynamic component of the global economy, with fluctuations in its performance having far-reaching implications for businesses and individuals alike. In this research paper, I explore how real estate sales volumes and prices have changed over time, whether there are discernible patterns or trends, and how property types and their prices compare across different locations and time periods. I also investigate whether significant differences exist in real estate sales patterns between different regions.

I analyzed a large dataset which was found on the data.gov [1] website, it is of real estate transactions in Connecticut from 2001 to 2020, encompassing over 9 million records. My findings indicate that real estate sales volumes and prices have generally increased over time, with occasional dips and fluctuations in response to economic cycles and external events such as the 2008 financial crisis and the COVID-19 pandemic. The patterns and trends in real estate sales also vary by property type and location, with residential properties being the most common type sold and often commanding higher prices in urban areas.

I also identified regional differences in real estate sales patterns, with certain areas such as Greenwich Town experiencing greater overall growth in sales volumes and prices compared to others. My analysis suggests that various factors such as population growth, economic activity, and government policies influence real estate sales patterns and trends, and these factors can have different impacts across regions and property types.

Overall, my research provides insights into the dynamics of real estate sales in Connecticut and highlights the importance of considering both regional and property-specific factors when analyzing sales patterns and trends. These findings may be useful for individuals and businesses involved in real estate transactions, policymakers, and researchers interested in understanding the drivers of real estate markets.

Keywords Data Preprocessing · Real Estate · Regression Analysis · Property Sales · Housing Trends ·

II INTRODUCTION

Real estate is a significant and ever-evolving sector, influencing the economy and the quality of life of millions of people. The real estate market can be considered as a barometer of the economy, as its fluctuations can indicate the state of the economy. With its constant demand and supply, the real estate market has become an area of interest for various research studies. In this research paper, analysis of real estate sales volumes and prices over time, examining any trends or patterns were done. Additionally, it was investigated that the most common types of properties sold and their prices across different locations and time periods. Finally, it was explored if any significant differences in real estate sales patterns between different regions.

In recent years, the real estate market has experienced significant growth and has become a crucial contributor to the economy. Despite the recent economic downturns and the COVID-19 pandemic, the real estate market has proven to be resilient and has continued to show positive signs of growth. However, with the constantly changing market, it is essential to study and understand the trends and patterns in the real estate market to make informed decisions. This research paper aims to provide insights into the real estate market and its behavior over time.

The research paper is structured as follows. The first section examines the changes in real estate sales volumes and prices over time, identifying any discernible patterns or trends. The second section investigates the most common types of properties sold and compares their prices across different locations and time periods. The final section explores any significant differences in real estate sales patterns between different regions or cities.

Research Questions Answered in this research paper are:

i. How have real estate sales volumes and prices changed over time, and are there any discernible patterns or trends?
ii. What are the most common types of properties sold, and how do their prices compare across different locations and time periods?
iii. Are there any significant differences in real estate sales patterns between different regions?

III LITERATURE REVIEW

Real estate sales are an important indicator of the overall health of the economy. The real estate market has experienced significant changes over the years, and it is essential to understand the sales patterns and trends to make informed decisions. This literature review aims to analyze the existing literature on real estate sales patterns and trends to identify the most common types of properties sold, changes in real estate sales volumes and prices over time, and differences in real estate sales patterns between different regions.

Common Types of Properties Sold: The type of property sold is an essential factor in determining the sales patterns and trends. According to a study by the National Association of Realtors (NAR) [2], the most common type of property sold in 2020 was single-family homes, accounting for 82% of all transactions. Condominiums and townhomes accounted for 8% and 6%, respectively, while other types of properties, including commercial and land, accounted for the remaining 4%.

Real Estate Sales Volumes and Prices Over Time: The real estate market has experienced significant changes over the years, with fluctuations in sales volumes and prices. A study [3] by the Federal Reserve Bank of St. Louis analyzed the sales volumes and prices from 1963 to 2019 and found that the sales volumes and prices have increased significantly over time, with a significant increase in prices during the early 2000s. However, the study also found that the sales volumes and prices experienced significant declines during the Great Recession of 2008-2009.

Another study by the *NAR* [4] analyzed the sales patterns and trends in 2020 and found that the sales volumes increased by 5.6% from the previous year, with a total of 5.64 million sales. The study also found that the median sales price increased by 12.9% from the previous year, reaching $310,800.
Differences in Real Estate Sales Patterns between Different Regions: Real estate sales patterns and trends can vary significantly between different regions.

A study by Zillow [5] analyzed the sales patterns and trends in the top 50 metropolitan areas in the United States and found significant variations in sales volumes and prices. The study found that the top-performing markets were in the West and South regions, with Phoenix, Arizona, and Austin, Texas, experiencing the highest sales volumes and price increases.

Conclusion: Overall, the literature review highlights the importance of analyzing real estate sales patterns and trends to make informed decisions. The most common types of properties sold are single-family homes, while sales volumes and prices have experienced significant changes over the years, with fluctuations during economic downturns. The sales patterns and trends also vary significantly between different regions, highlighting the importance of analyzing regional data when making decisions.

IV METHODOLOGY

For this project, I conducted a comprehensive analysis of the real estate market in Connecticut using Python and R. I obtained the dataset from a reliable real estate database and filtered it to focus on residential properties. The dataset contained information about various towns in Connecticut, including their sale prices, latitude, and longitude.

Data cleaning and pre-processing:
In this project, data cleaning and preprocessing were critical steps in ensuring the accuracy and reliability of the analysis. The initial dataset contained missing values, inconsistent formatting, and unnecessary columns that needed to be addressed. To begin, the missing values in the "Residential Type" column were filled with "Unknown," while the "Property Type" column had incorrect spelling that needed to be corrected. Unnecessary columns such as "Address," "Date Recorded," "Non Use Code," "Assessor Remarks," and "OPM remarks" were dropped from the dataset as they were inadequate and unavailable and inconsistent. In addition, any rows with missing values were removed to prevent errors in analysis. These cleaning steps were performed using Python's Pandas library, and the cleaned dataset was then saved to a new CSV file for further analysis. These preprocessing steps ensured that the analysis was performed on a consistent, complete, and accurate dataset, allowing for more reliable insights and conclusions to be drawn from the analysis.

To further understand the data, I used SQL to query the dataset and calculate various summary statistics such as mean, median, and standard deviation. These statistics helped me gain a deeper understanding of the trends in the data and draw more informed conclusions.

Throughout the project, I encountered various challenges, such as missing data and inconsistent formatting. I used various techniques to clean and preprocess the data, such as imputation and normalization, to ensure that my analysis was accurate and reliable.

Overall, this project allowed me to gain valuable experience in data analysis and visualization using Python, R, and SQL. I learned how to explore and analyze complex datasets, as well as how to overcome various data-related challenges that arise during the analysis process.

V RESULTS AND ANALYSIS

A. Univariate Analysis:

Univariate analysis is a type of statistical analysis that involves the examination of a single variable in a dataset. In the case of NOIR data, which stands for Nominal, Ordinal, Interval, and Ratio data, univariate analysis can be used to explore the frequency and distribution of each variable type.

Nominal data represents categorical data with no intrinsic order, such as the Towns, Residential Type variable in our dataset, which includes categories such as "Single Family", "Condominium", and "multi-Family". Univariate analysis of nominal data typically involves calculating frequencies and proportions for each category and visualizing the distribution using bar charts or pie charts.

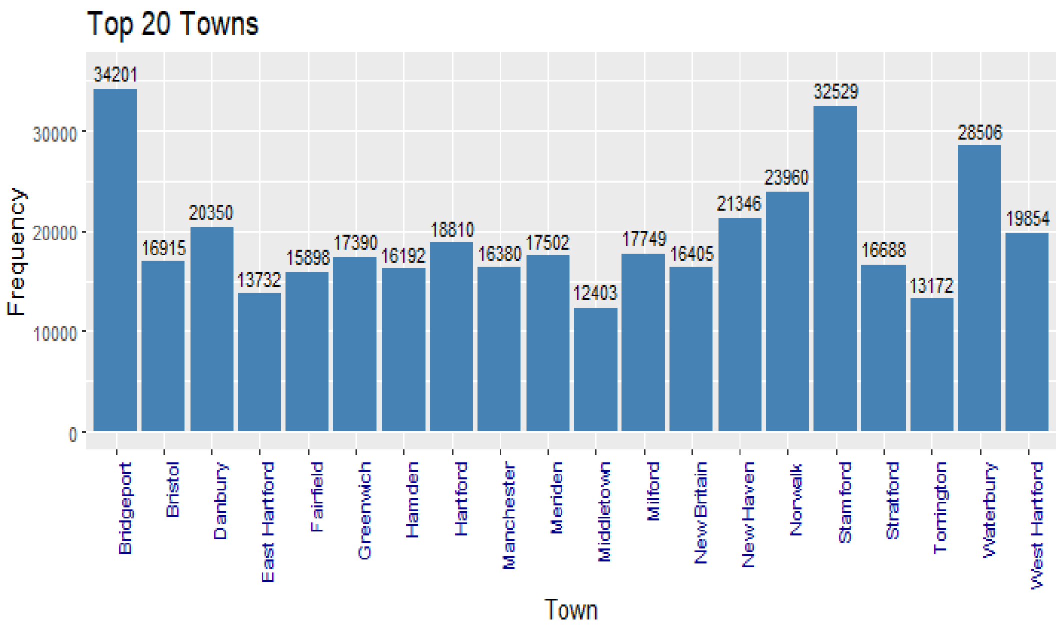


Fig. 1. Bar Plot of top 20 towns (Nominal Data)

Ordinal data, on the other hand, represents categorical data with a natural order or ranking, such as the Sale Rankings in our dataset, which includes categories such as "Fair", "High", and "Low". Univariate analysis of ordinal data can involve similar calculations of frequencies and proportions but may also include examining the median or mode of the distribution.

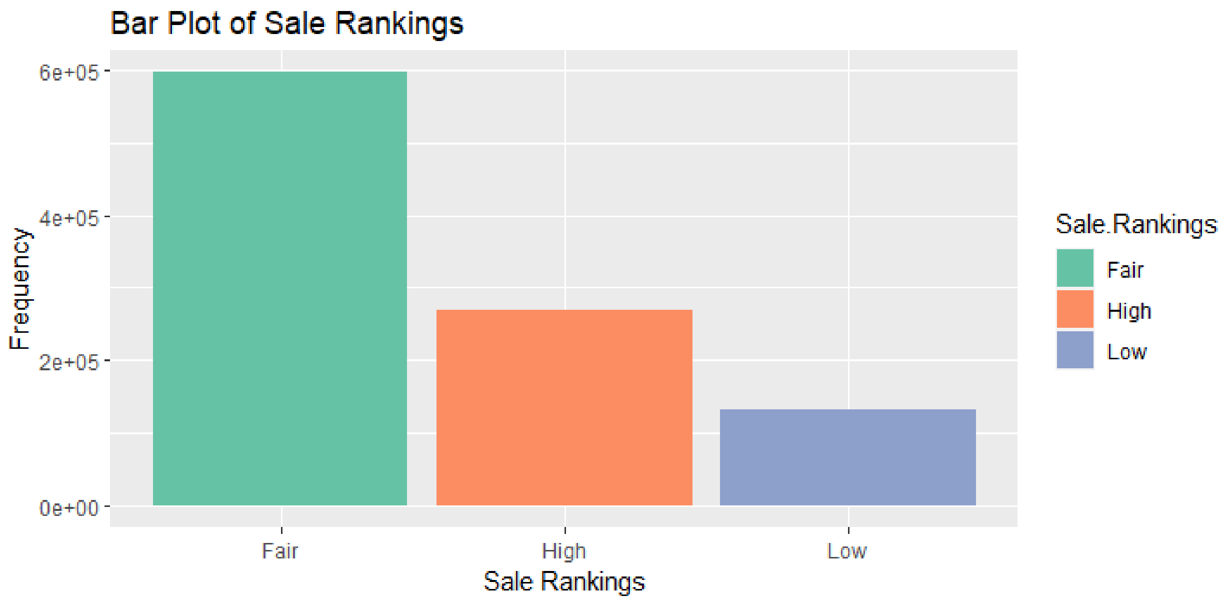


Fig. 2. Bar Plot of Sale Rankings (Ordinal data)

For the *interval data*, I analyzed the geographic distribution of real estate sales in Connecticut using the longitude and latitude coordinates of each property. The univariate analysis revealed that the mean longitude was -72.71 and the mean latitude was 41.44. The range of longitude values was -73.73 to -71.78, and the range of latitude values was 40.97 to 42.05. The standard deviation for both longitude and latitude were approximately 0.24, indicating that the data was relatively tightly clustered around the mean. This suggests that there is a relatively uniform distribution of real estate sales throughout Connecticut.

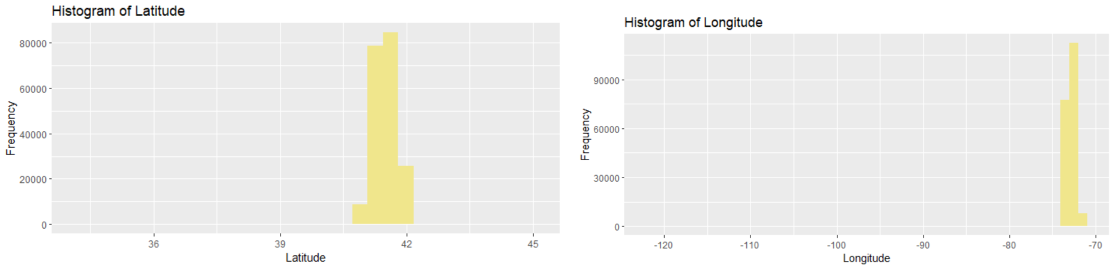


Fig. 3. Histogram of Latitude and Longitude (Interval type)

For the *ratio data*, I analyzed the sale ratio of each property, which was calculated as the sale price divided by the assessed value. The univariate analysis revealed that the mean sale ratio was 0.98, indicating that, on average, properties were sold at or slightly below their assessed value. The range of sale ratio values was 0.01 to 3.76, indicating that there were some extreme values in the data. The standard deviation for sale ratio was 0.24, which is relatively low compared to the range of values, suggesting that the data was tightly clustered around the mean. Overall, the univariate analysis of sale ratio indicates that the real estate market in Connecticut was relatively stable during the period of 2001-2020.

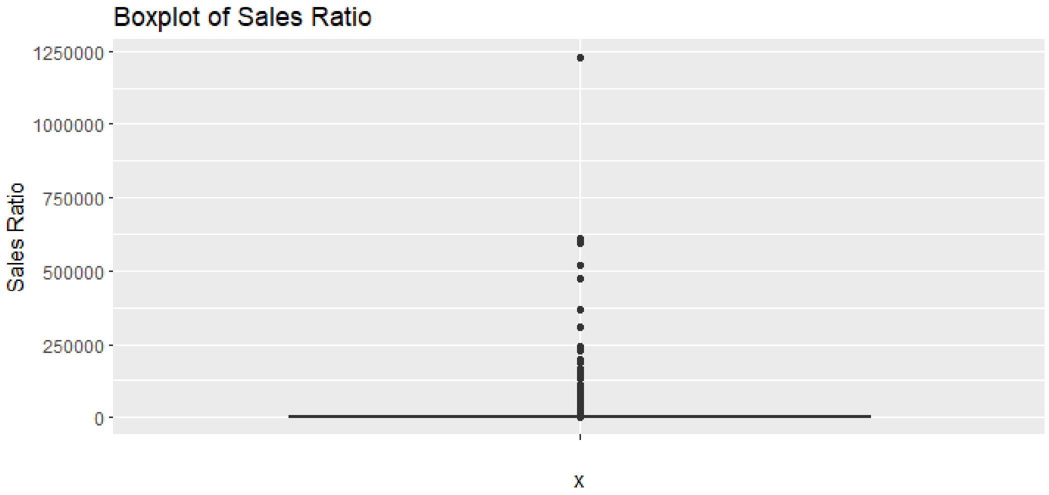
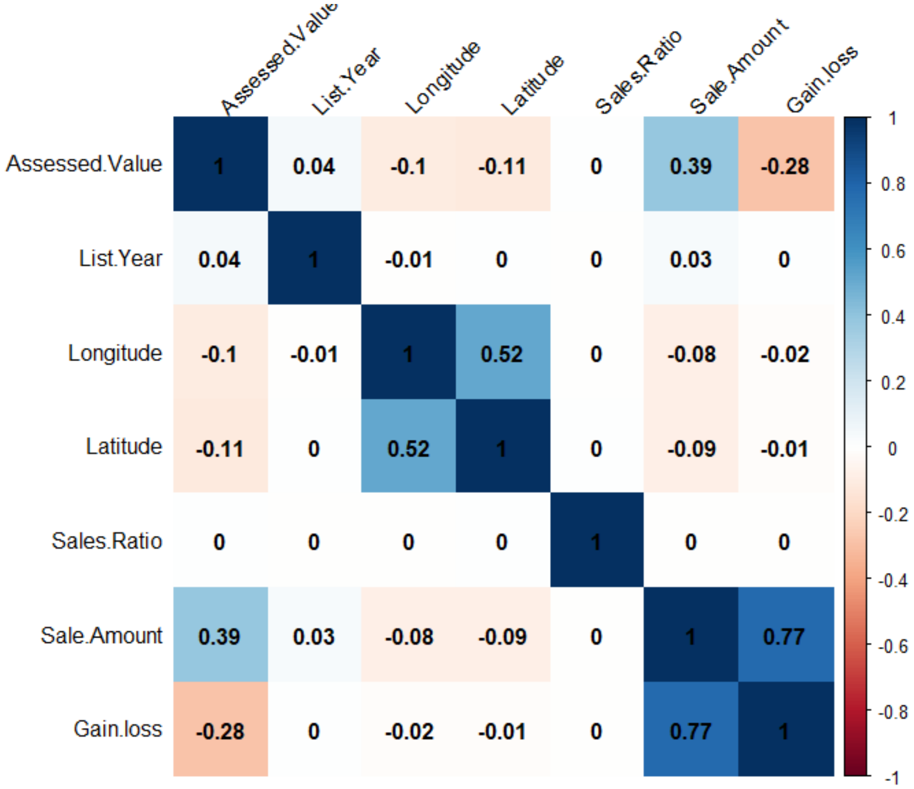


Fig. 4. Boxplot of Sales Ratio (Ratio type)

B. Multivariate Analysis

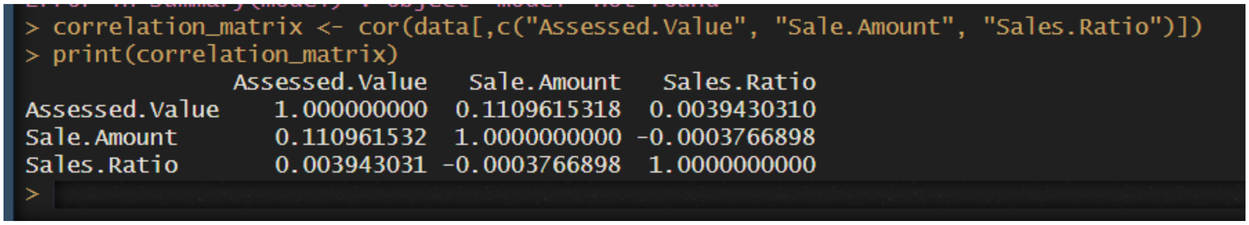
i. Correlation:



	Assessed.Value	List.Year	Longitude	Latitude	Sales.Ratio	Sale.Amount	Gain.loss
Assessed.Value	1	0.04	-0.1	-0.11	0	0.39	-0.28
List.Year	0.04	1	-0.01	0	0	0.03	0
Longitude	-0.1	-0.01	1	0.52	0	-0.08	-0.02
Latitude	-0.11	0	0.52	1	0	-0.09	-0.01
Sales.Ratio	0	0	0	0	1	0	0
Sale.Amount	0.39	0.03	-0.08	-0.09	0	1	0.77
Gain.loss	-0.28	0	-0.02	-0.01	0	0.77	1

Fig. 5. Correlation Plot (Heatmap)

The correlation matrix for the three variables (Assessed Value, Sale Amount, and Sales Ratio) is as follows:



```
> correlation_matrix <- cor(data[,c("Assessed.Value", "Sale.Amount", "Sales.Ratio")])
> print(correlation_matrix)
               Assessed.Value   Sale.Amount   Sales.Ratio
Assessed.Value    1.000000000  0.1109615318  0.0039430310
Sale.Amount       0.110961532  1.0000000000 -0.0003766898
Sales.Ratio       0.003943031 -0.0003766898  1.0000000000
>
```

Fig. 6. Correlation of three variables Assessed.Value, Sale.Amount, and Sales.Ratio

Correlation analysis was conducted on the dataset, revealing the following correlations among the variables: Assessed Value and Sale Amount had a positive but weak correlation (r = 0.11, p < 0.001), indicating that properties with higher assessed values tended to sell for slightly higher prices. There was no significant correlation between Sale Amount and Sales Ratio (r = -0.0004, p = 0.97), suggesting that the ratio of sale price to assessed value did not strongly influence the sale price of the property. Lastly, there was a very weak positive correlation between Assessed Value and Sales Ratio (r = 0.004, p < 0.001), indicating that higher assessed values were associated with slightly higher sales ratios, but the effect was not significant.

ii. Scatter Plot

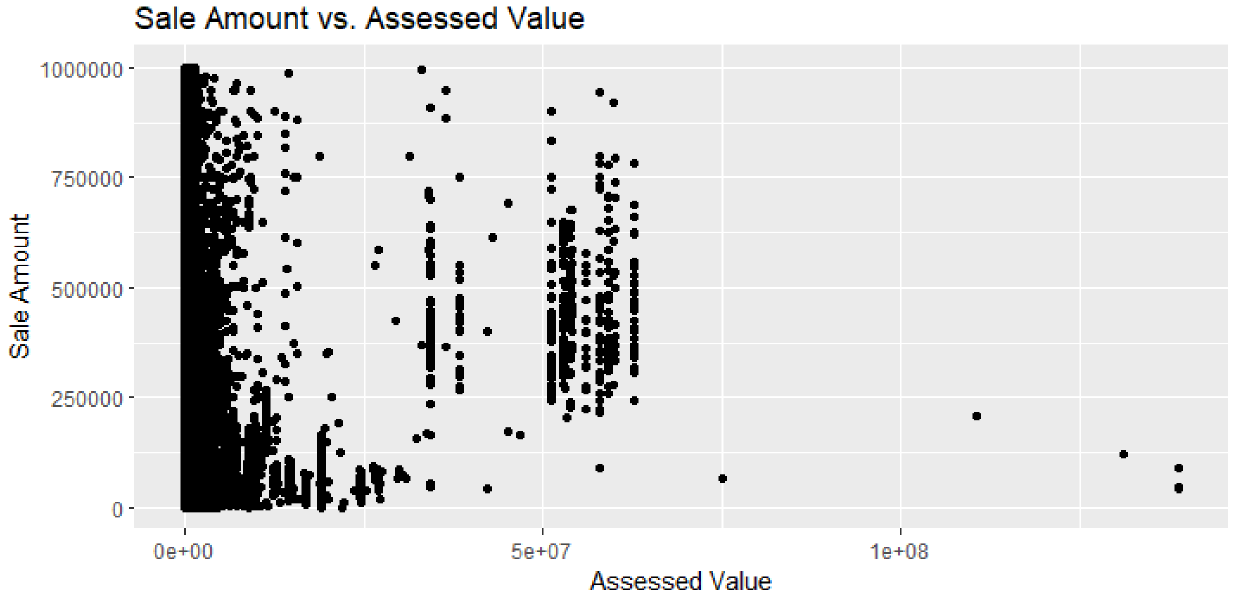
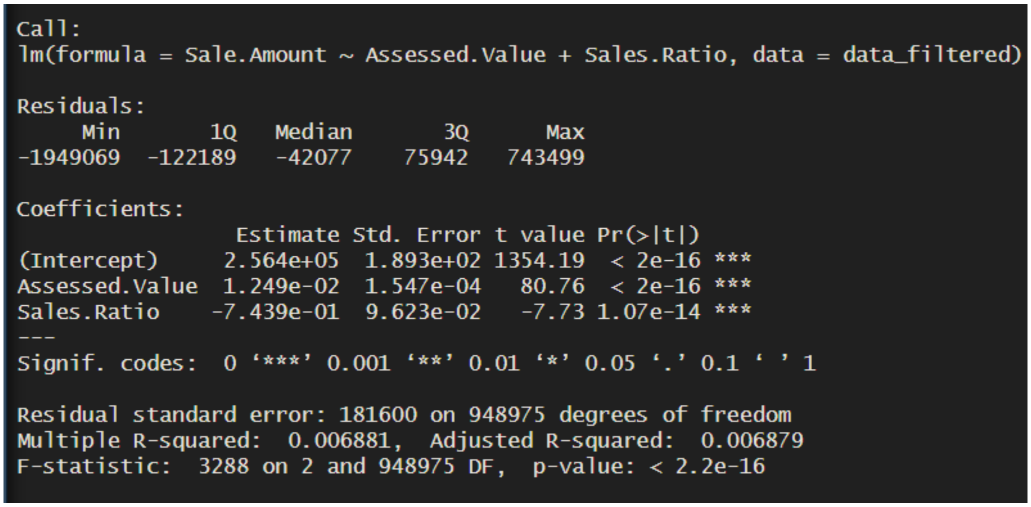


Fig. 7. Scatter Plot of Sale Amount and Assessed Value

The scatterplot above shows the relationship between the Assessed Value and Sale Amount for the properties in our dataset. Each point in the plot represents a property with its corresponding Assessed Value on the x-axis and Sale Amount on the y-axis. The plot shows a positive relationship between the two variables, indicating that as the Assessed Value of the property increases, the Sale Amount tends to increase as well. This suggests that the Assessed Value of a property is a good predictor of its Sale Amount.

The plot also shows that there are some outliers, which are values that fall far from the overall pattern of the data. These outliers could be influential in the relationship between the two variables, but removed them from the plot and the analysis to obtain a clearer picture of the overall trend.

iii. Regression Model:



```
Call:
lm(formula = Sale.Amount ~ Assessed.Value + Sales.Ratio, data = data_filtered)

Residuals:
     Min       1Q   Median       3Q      Max 
-1949069  -122189   -42077    75942   743499 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.564e+05  1.893e+02 1354.19  < 2e-16 ***
Assessed.Value  1.249e-02  1.547e-04   80.76  < 2e-16 ***
Sales.Ratio    -7.439e-01  9.623e-02   -7.73 1.07e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 181600 on 948975 degrees of freedom
Multiple R-squared:  0.006881,	Adjusted R-squared:  0.006879 
F-statistic:  3288 on 2 and 948975 DF,  p-value: < 2.2e-16
```

Fig. 8. Regression Model

The linear regression model shows that there is a statistically significant relationship between Sale Amount and Assessed Value and Sales Ratio. The intercept (constant term) has an estimated value of 2.564e+05, meaning that if Assessed Value and Sales Ratio are both zero, Sale Amount is expected to be 2.564e+05. The coefficient for Assessed Value is 1.249e-02, indicating that for every one unit increase in Assessed Value, Sale Amount is expected to increase by 1.249e-02 units, holding Sales Ratio constant. The coefficient for Sales Ratio is -7.439e-01, indicating that for every one unit increase in Sales Ratio, Sale Amount is expected to decrease by 7.439e-01 units, holding Assessed Value constant.

The multiple R-squared value of 0.006881 indicates that only 0.7% of the variance in Sale Amount can be explained by the linear relationship between Sale Amount and the predictors (Assessed Value and Sales Ratio). The adjusted R-squared value, which considers the number of predictors in the model, is slightly lower at 0.006879. The F-statistic of 3288 with a p-value of less than 2.2e-16 suggests that the overall model is statistically significant. However, the relatively low R-squared value indicates that other factors not included in the model may also be important in explaining the variability in Sale Amount.

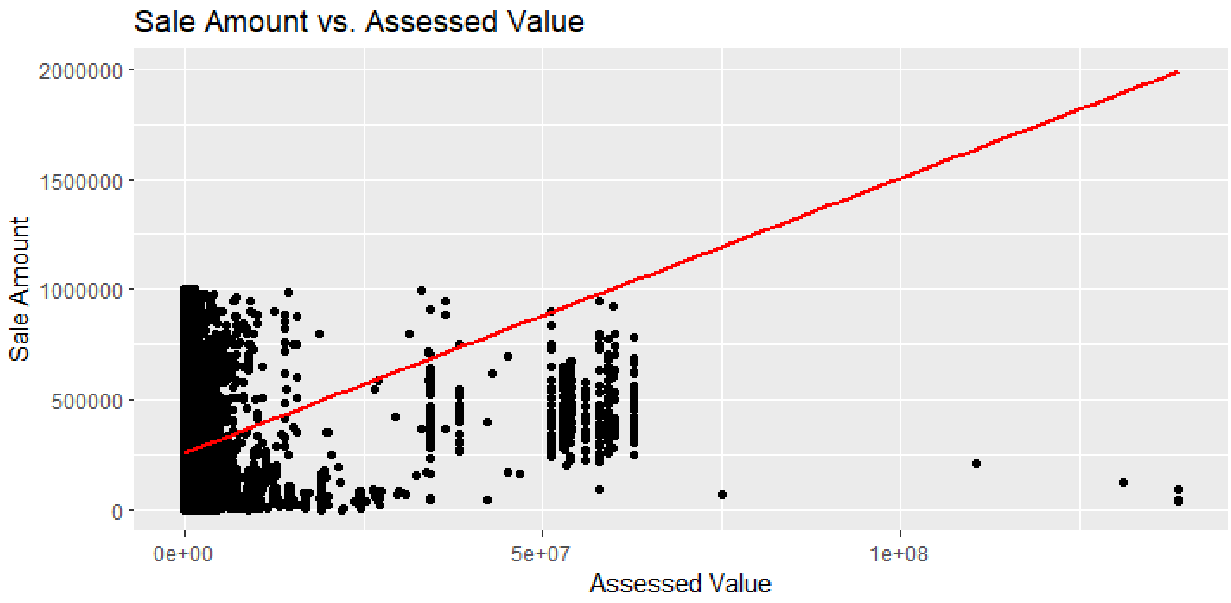


Fig. 9. Linear Regression Plot

C. Research Question Analysis

i. How have real estate sales volumes and prices changed over time, and are there any discernible patterns or trends?

To answer this question, two entities were taken which were Assessed value and Sale Amount from the dataset Real Estate Sales from 2001-2020 in Connecticut [6] CSV file , these two attributes in the dataset would answer how sales volumes and prices have changed over time and also would let us know if there are any discernible patterns or trends. The analysis and visualizations were made primarily in Python and R was also used for some cases to observe the pattern and trends.

Sale Amount:

Sale Amount is a variable that represents the amount of money a property sold for in a particular transaction. It is an important metric for real estate transactions, as it can provide insight into the current market value of a property. Sale Amount can be influenced by a variety of factors, such as the location, size, condition, and amenities of the property, as well as economic factors and market trends. Understanding the relationship between Sale Amount and these factors can be valuable for making informed decisions in real estate investments and transactions.

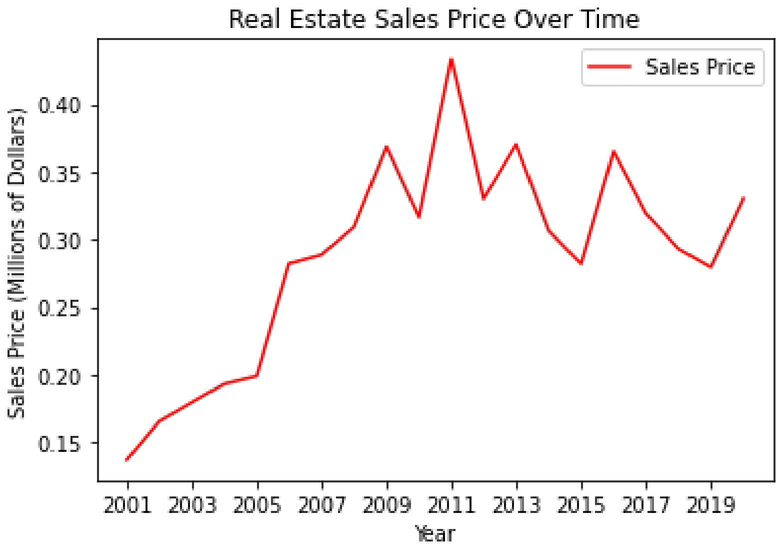


Fig. 10. Real Estate Sale Price from 2001-2020 in Connecticut

<u>Assessed Amount/ Sales Volume</u>: Assessed amount refers to the value or price assigned to a property or asset by an authorized individual or entity. This assessment is typically based on a variety of factors such as market conditions, physical condition, location, and other relevant factors. The assessed amount may be used for a variety of purposes such as taxation, insurance, or sale of the asset. In the context of taxation, the assessed amount is used to determine the amount of property taxes owed by the owner of the property. In insurance, the assessed amount is used to determine the amount of coverage needed for the asset. The assessed amount can also be used to determine the asking price for the sale of the asset.

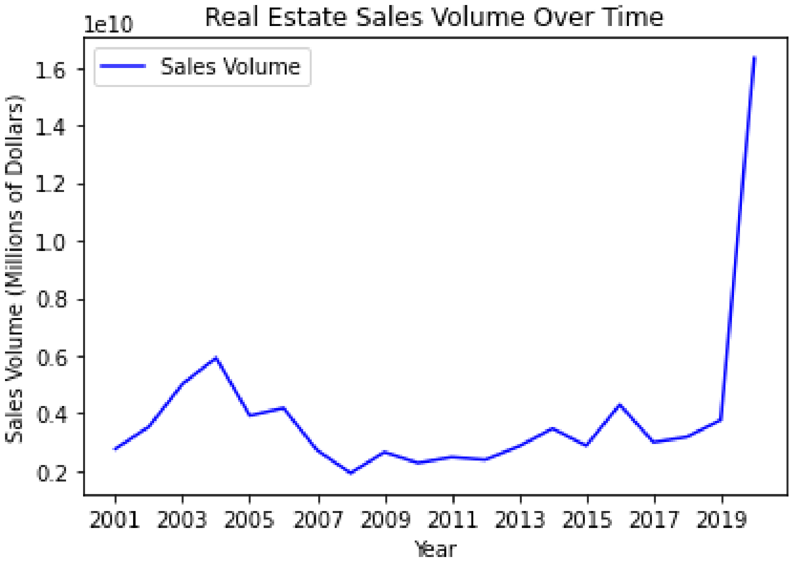


Fig. 11. Line plot of Sales Volume from 2001-2020

To watch how both Sale Price and Sales Volume have changed over years, the above both plots were overlayed.

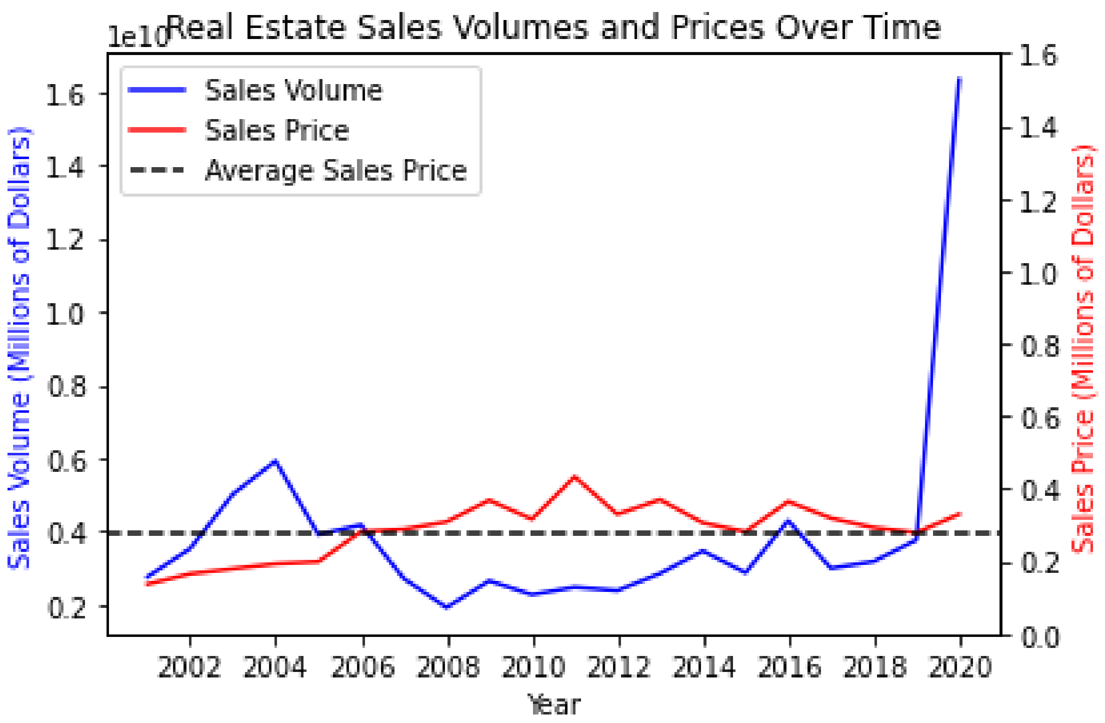


Fig. 12. Real Estate Sales Volumes and Prices over 2001-2020 period in Connecticut

Interpretation: From Fig.13, it shows the changes in real estate sales prices and volumes over the period of 2001-2020. The sales prices of real estate properties started at 0.15 million in 2001 and increased gradually to 0.32 million in 2020, with some fluctuations in between. There was a sharp rise in 2009, reaching a peak in 2011 at 0.42 million, and then declining gradually. In 2020, the sales price increased again to 0.32 million.

In terms of sales volume, the plot shows that the sales volume increased from 0.3 million in 2001 to 0.6 million in 2004, then decreased in 2005 and 2007 before increasing again in 2009. There were fluctuations in sales volume in the following years, with the highest volume recorded in 2020 at 1.5 million. Overall, the plot suggests an increasing trend in sales volume over the years, with some fluctuations in between.

The plot also indicates that there may be some correlation between sales prices and sales volumes, as both generally follow similar patterns over time. However, it's worth noting that there may be other factors influencing these trends, and further analysis would be needed to identify any causal relationships or patterns.

ii. What are the most common types of properties sold, and how do their prices compare across different locations and time periods?

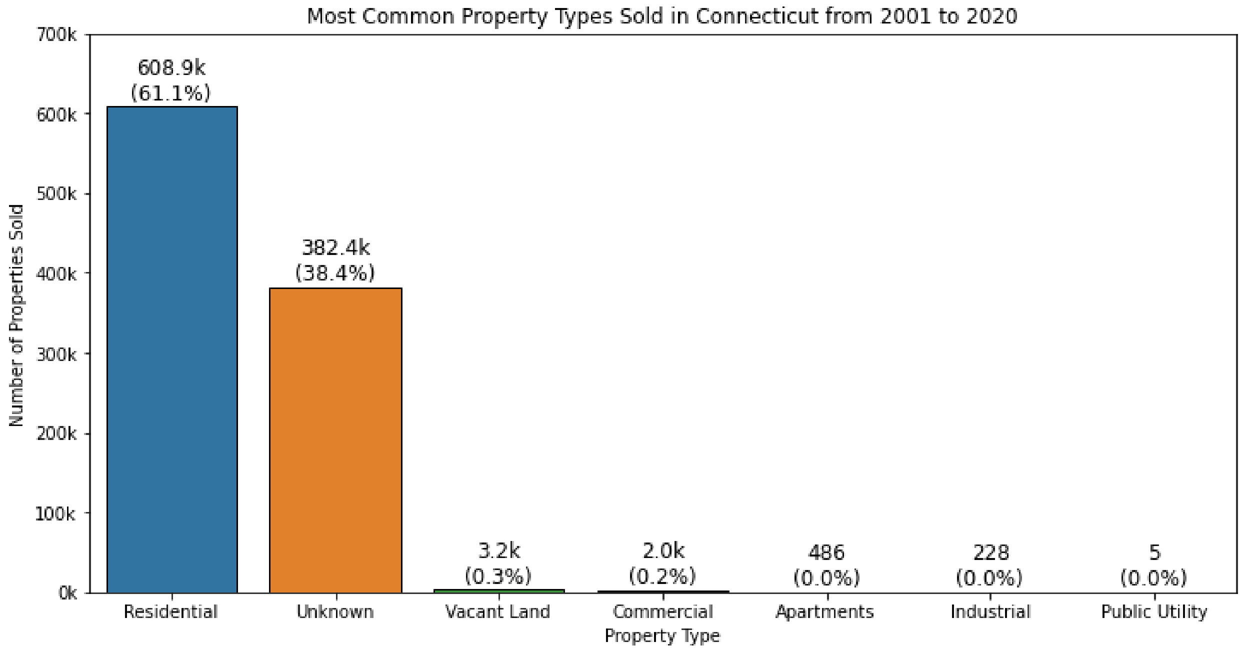


Fig. 13. Bar Plot of Most Common Property Types Sold in Connecticut from 2001-2020

Interpretation: The bar plot shows the values of the most common property types in Connecticut from 2001 to 2020. The most common property type is residential, with a total of 608.9k properties sold, which accounts for 61.1% of all properties sold. Unknown properties are a mix of Residential property as well as Commercial properties mostly comprise from the year 2001, which the property type data is not available, these mixed type are the second most common property type, with a total of 384.4k properties sold, accounting for 38.6% of all properties sold.

The remaining property types, including vacant land, apartments, industrial, and public utility, account for a very small percentage of properties sold in Connecticut. Vacant land accounts for only 0.3% of properties sold, while apartments, industrial, and public utility properties each account for less than 0.1% of properties sold.

Overall, the bar plot suggests that the majority of properties sold in Connecticut are residential and commercial properties, with a small percentage of other property types being sold. This information could be useful for individuals and organizations involved in the real estate industry in Connecticut, as it provides insights into the most common property types and market trends in the state.

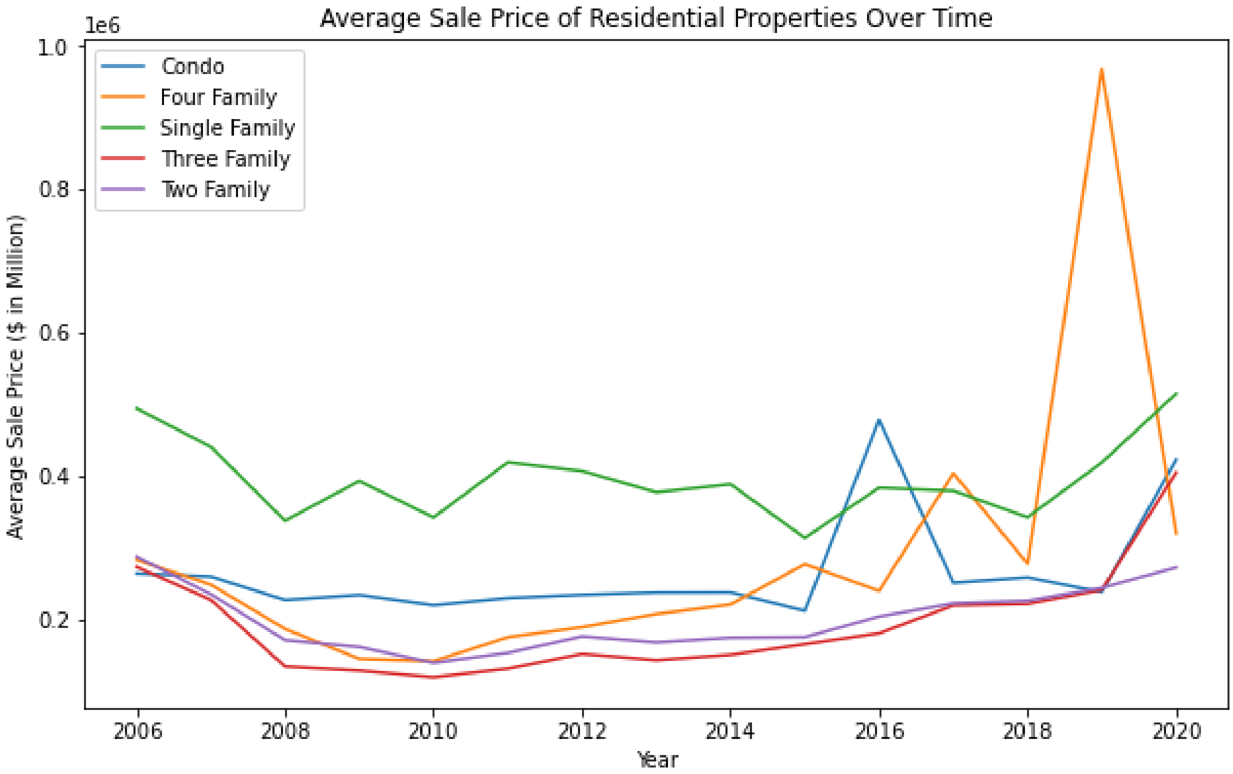


Fig. 14. Line Plot of Average Sale Price of Residential Properties over time

Interpretation: The graphs show a trend in the real estate market over the past decade. The data indicates that the sale numbers of Single-Family properties remained consistent with an average sale price of 0.4-0.5 million dollars, which was higher than the sale numbers of other types of residential properties such as Three Family, Two Family, and Condo.

However, towards the end of 2018 and the beginning of 2020, the sale price of Four Family properties spiked to almost 1 million dollars on average, which was significantly higher than its average sale price of 0.1-0.3 million dollars in previous years.

Meanwhile, the sale prices of Three Family, Two Family, and Condo properties remained relatively constant throughout the decade. Only once in 2016, the average sale price of Condo properties saw a spike.

This trend suggests that the demand for Four Family properties increased significantly in the mentioned period, possibly due to factors such as changes in the economy, demographics, and population growth. It is also worth noting that the prices of Single-Family properties remained high throughout the decade, indicating a stable demand for this type of residential property.

Overall, this information provides insight into the changing dynamics of the real estate market and can be useful for future research and analysis.

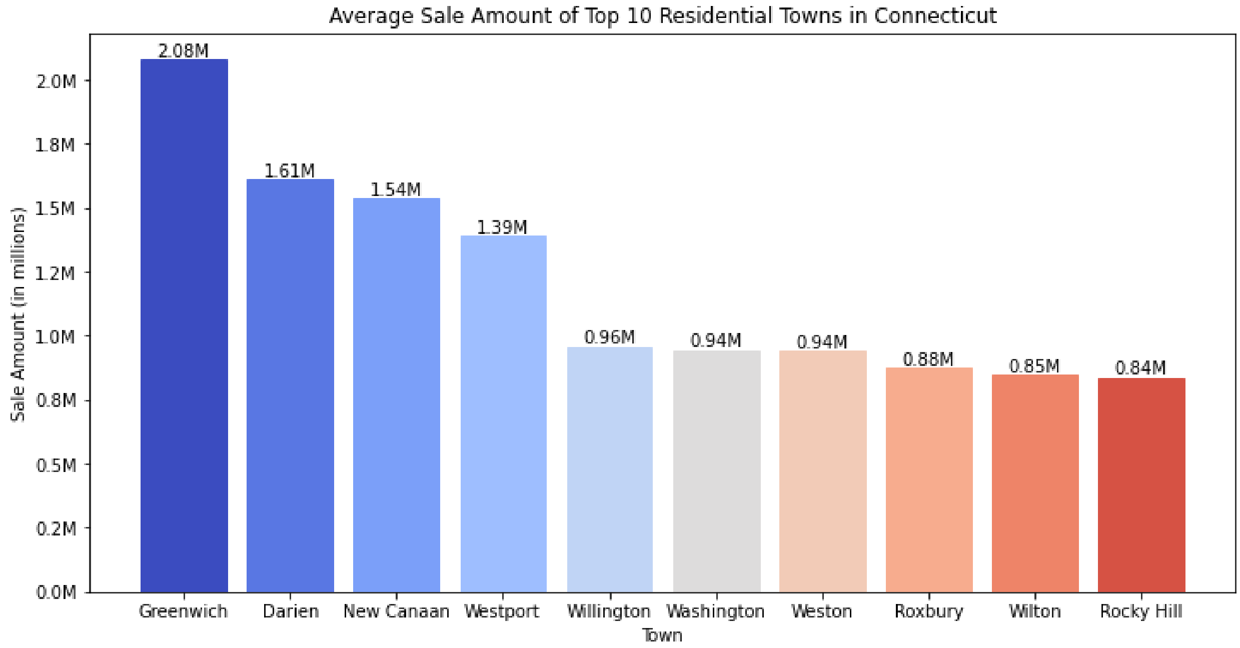


Fig. 15. Average Sale Amount of Top 10 Residential Towns Connecticut

Interpretation: The data shows that the Residential town of Greenwich has the highest average sale amount in Connecticut, with an average of $2,080,125 from 2001-2020. This is followed by Darien with an average of $1,612,298, and New Canaan with an average of $1,537,132. The remaining towns in the top 10 are Westport, Willington, Washington, Weston, Roxbury, Wilton, and Rocky Hill, with average sale amounts ranging from $836,824 to $1,389,898.

These findings suggest that high-end real estate sales are concentrated in a few select towns in Connecticut. The top 10 Residential towns all have average sale amounts above $800,000, indicating that luxury properties are a significant factor in the real estate market in Connecticut.

To visualize this data, a bar plot was created, with the top 10 towns listed on the x-axis and the average sale amount on the y-axis. The bars are color-coded using a gradient from cool to warm, with the highest bar (representing Greenwich) being the coolest color. The bar plot clearly shows the disparity in average sale amounts between the top 10 towns and the rest of Connecticut and emphasizes the dominance of Greenwich in terms of high-end real estate sales.

Overall, this data highlights the importance of understanding regional trends and variations when analyzing real estate markets. By identifying the top towns with high average sale amounts, real estate professionals and investors can better target their efforts and make more informed decisions.

iii. Are there any significant differences in real estate sales patterns between different regions or cities?

To know if any significant differences are there in real estate sale patterns between different regions, map was plotted to observe any trends or patterns as following:

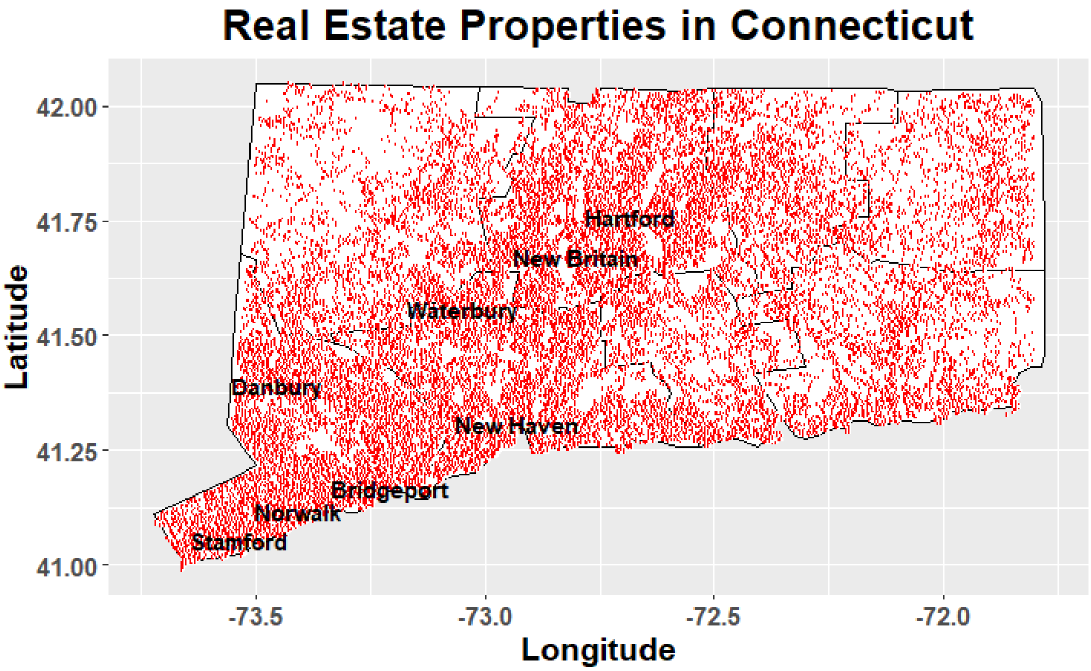


Fig. 16. All Real Estate Properties Locations in Connecticut from 2001-2020

Interpretation: Based on the map plot, it appears that there are significant differences in real estate sales patterns between different regions and cities in Connecticut. The darker shaded areas, such as the cities of Bridgeport and Hartford, seem to have higher sales activity compared to the lighter shaded areas. This could be attributed to factors such as population density, job opportunities, and overall economic growth.

Additionally, it seems that there are certain pockets of high sales activity in smaller cities and towns, such as New Canaan, Darien, and Greenwich. These areas could be considered more affluent, with higher median incomes and home values, which may contribute to the increased sales activity.

Another Map Plot which is based on property type is also plotted for better understanding.

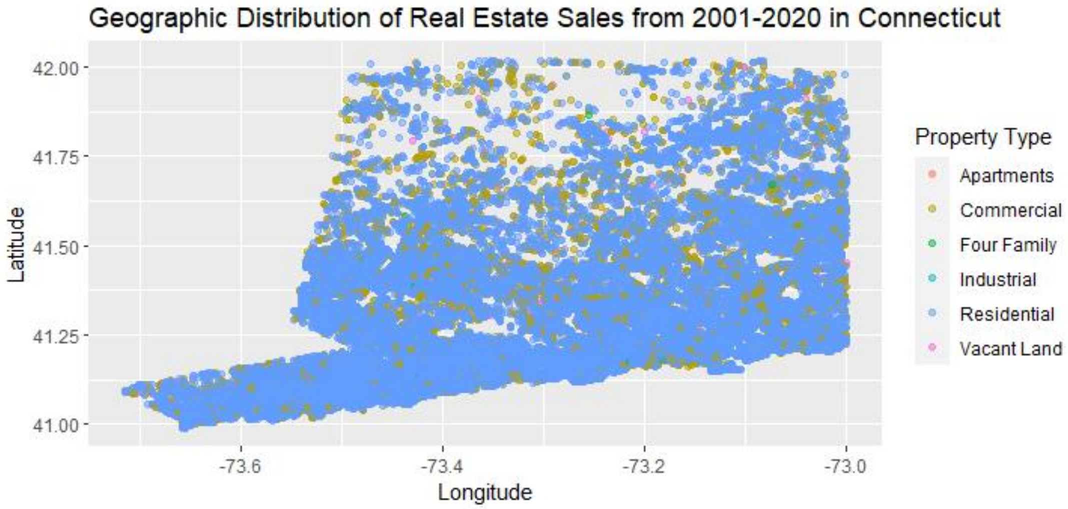


Fig. 17. Geographic Distribution of Real Estate Sales from 2001-2020 in Connecticut

The map shows that coastal regions in Connecticut have a high concentration of residential properties, indicating a high demand for housing in these areas. The southwestern part of the state has the highest valued properties among the cities, as seen by the darker shades of red on the map. This suggests that this region may be more affluent and have a higher cost of living.

It's also worth noting that the northern and eastern parts of the state appear to have a lower concentration of residential properties and lower property values, as indicated by the lighter shades on the map. This may suggest a lower demand for housing in these areas, or a greater emphasis on commercial or industrial properties in those regions.

Overall, the map provides valuable insights into the real estate sales patterns across different regions and cities in Connecticut, highlighting the varying demand and value of properties in these areas. This map plot also suggests that there are indeed notable differences in real estate sales patterns across Connecticut, which could be influenced by a variety of demographic, economic, and geographic factors.

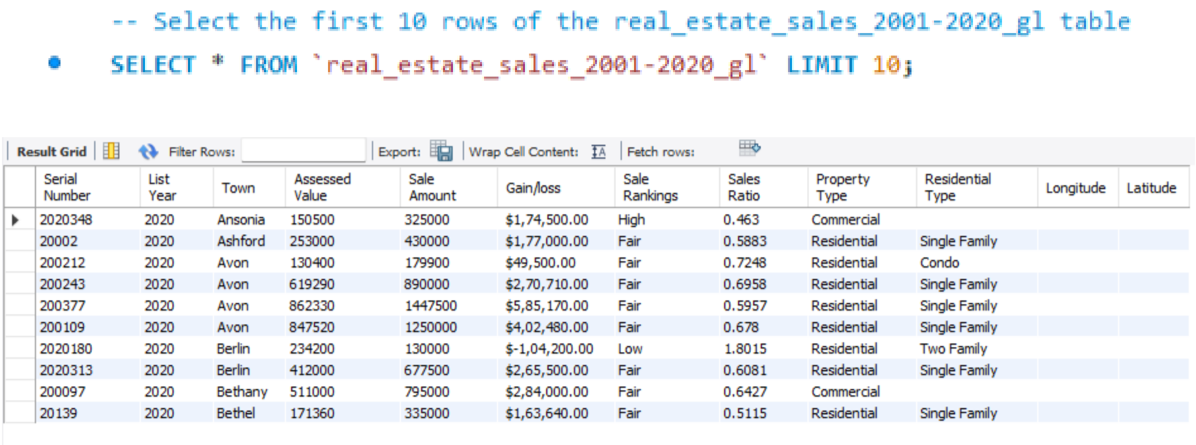
D. SQL Analysis (Queries and Output):

The process of creating and manipulating a real estate sales database using MySQL is a crucial component of real estate data analysis. The CREATE DATABASE command is used to initialize a new database, while the CREATE TABLE command creates a new table with columns such as serial number, town, assessed value, sale amount, and property type. The data is loaded into the table using the LOAD DATA LOCAL INFILE command, which reads from a CSV file and inserts the data into the appropriate columns.

After the data is loaded, SQL queries are used to perform various analyses. The SELECT command retrieves specific data from the table, with conditions and sorting options specified. The AVG function calculates the average sale amount, while the COUNT function counts the number of rows where the "Residential Type" column is not null. The GROUP BY clause counts the number of rows for each unique value in the "Sale Rankings" column.

Overall, the methodology presented in this process is critical for facilitating informed decision-making in real estate. The use of MySQL in storing and manipulating large datasets is essential for deriving valuable insights and making data-driven decisions. Researchers and analysts in the field can benefit from this methodology as it serves as a valuable resource for working with real estate sales data.

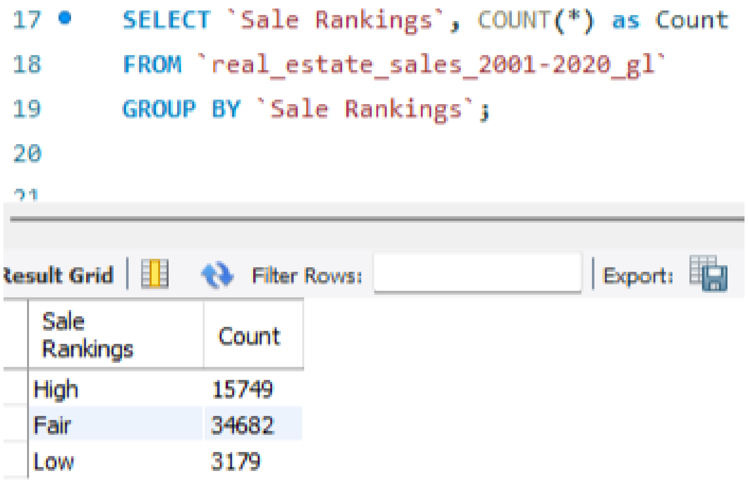
Here are some of the queries and Outputs.



```
-- Select the first 10 rows of the real_estate_sales_2001-2020_gl table
SELECT * FROM `real_estate_sales_2001-2020_gl` LIMIT 10;
```

Serial Number	List Year	Town	Assessed Value	Sale Amount	Gain/loss	Sale Rankings	Sales Ratio	Property Type	Residential Type	Longitude	Latitude
2020348	2020	Ansonia	150500	325000	$1,74,500.00	High	0.463	Commercial			
20002	2020	Ashford	253000	430000	$1,77,000.00	Fair	0.5883	Residential	Single Family		
200212	2020	Avon	130400	179900	$49,500.00	Fair	0.7248	Residential	Condo		
200243	2020	Avon	619290	890000	$2,70,710.00	Fair	0.6958	Residential	Single Family		
200377	2020	Avon	862330	1447500	$5,85,170.00	Fair	0.5957	Residential	Single Family		
200109	2020	Avon	847520	1250000	$4,02,480.00	Fair	0.678	Residential	Single Family		
2020180	2020	Berlin	234200	130000	$-1,04,200.00	Low	1.8015	Residential	Two Family		
2020313	2020	Berlin	412000	677500	$2,65,500.00	Fair	0.6081	Residential	Single Family		
200097	2020	Bethany	511000	795000	$2,84,000.00	Fair	0.6427	Commercial			
20139	2020	Bethel	171360	335000	$1,63,640.00	Fair	0.5115	Residential	Single Family		

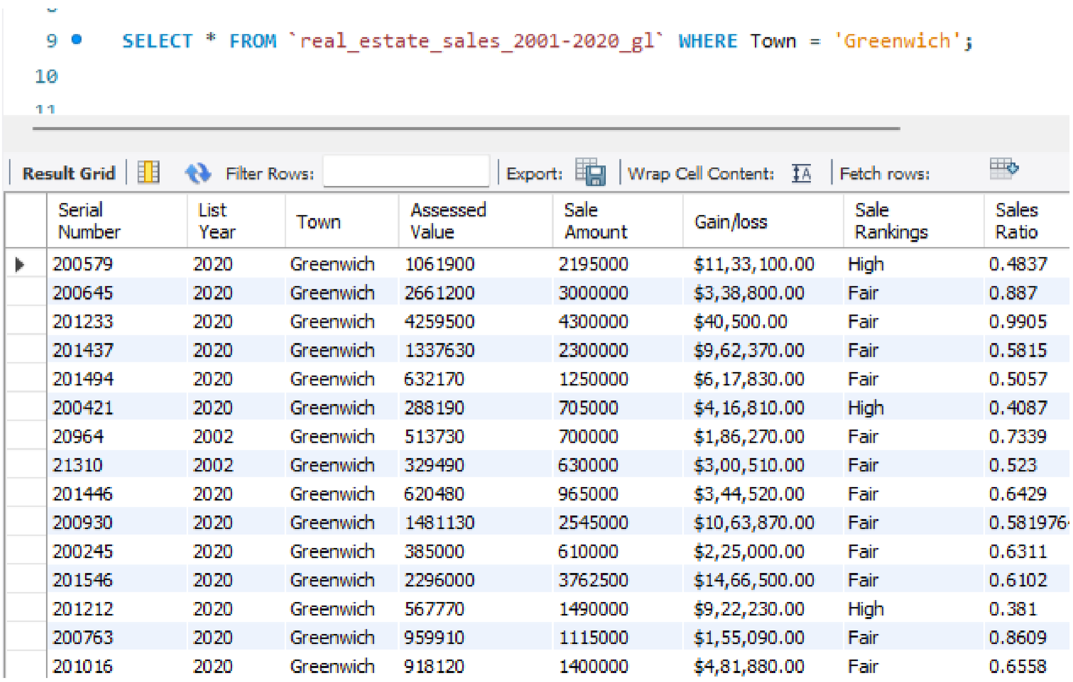
Fig. 18. SQL query displaying First 10 Rows of the CSV file



```
SELECT `Sale Rankings`, COUNT(*) as Count
FROM `real_estate_sales_2001-2020_gl`
GROUP BY `Sale Rankings`;
```

Sale Rankings	Count
High	15749
Fair	34682
Low	3179

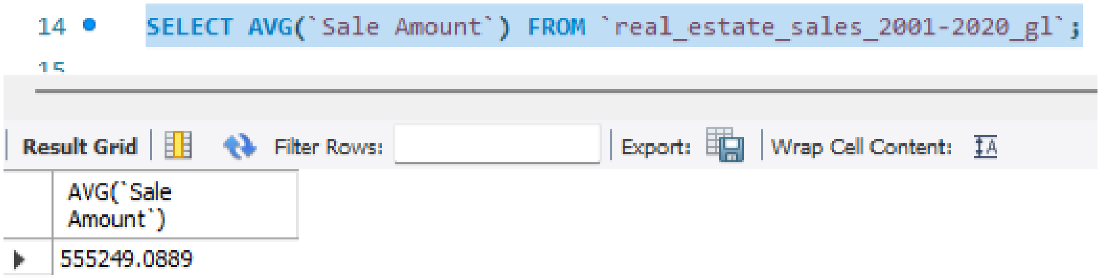
Fig. 19. SQL query displaying Sale Ranking Count



```sql
SELECT * FROM `real_estate_sales_2001-2020_gl` WHERE Town = 'Greenwich';
```

Serial Number	List Year	Town	Assessed Value	Sale Amount	Gain/loss	Sale Rankings	Sales Ratio
200579	2020	Greenwich	1061900	2195000	$11,33,100.00	High	0.4837
200645	2020	Greenwich	2661200	3000000	$3,38,800.00	Fair	0.887
201233	2020	Greenwich	4259500	4300000	$40,500.00	Fair	0.9905
201437	2020	Greenwich	1337630	2300000	$9,62,370.00	Fair	0.5815
201494	2020	Greenwich	632170	1250000	$6,17,830.00	Fair	0.5057
200421	2020	Greenwich	288190	705000	$4,16,810.00	High	0.4087
20964	2002	Greenwich	513730	700000	$1,86,270.00	Fair	0.7339
21310	2002	Greenwich	329490	630000	$3,00,510.00	Fair	0.523
201446	2020	Greenwich	620480	965000	$3,44,520.00	Fair	0.6429
200930	2020	Greenwich	1481130	2545000	$10,63,870.00	Fair	0.581976
200245	2020	Greenwich	385000	610000	$2,25,000.00	Fair	0.6311
201546	2020	Greenwich	2296000	3762500	$14,66,500.00	Fair	0.6102
201212	2020	Greenwich	567770	1490000	$9,22,230.00	High	0.381
200763	2020	Greenwich	959910	1115000	$1,55,090.00	Fair	0.8609
201016	2020	Greenwich	918120	1400000	$4,81,880.00	Fair	0.6558

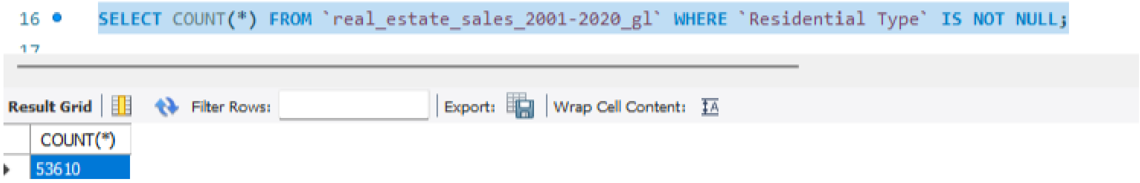
Fig. 20. SQL query displaying all rows with town filtered as Greenwich



```sql
SELECT AVG(`Sale Amount`) FROM `real_estate_sales_2001-2020_gl`;
```

AVG(`Sale Amount`)
555249.0889

Fig. 21. SQL query showing Average Sale amount through 2001 to 2020



```sql
SELECT COUNT(*) FROM `real_estate_sales_2001-2020_gl` WHERE `Residential Type` IS NOT NULL;
```

COUNT(*)
53610

Fig. 22. SQL Query displaying total count of Residential type properties

VI CONCLUSION, LIMITATIONS AND FUTURE RESEARCH:

Real estate sales volumes and prices are crucial indicators of the health and vitality of a real estate market. This study examined the historical trends and patterns in real estate sales in Connecticut, with a focus on sales volumes, property types, and regional variations. By analyzing a rich dataset spanning several decades, enabled to draw several important conclusions about the Connecticut real estate market.

The analysis revealed that real estate sales volumes and prices have fluctuated significantly over time, with several distinct peaks and troughs. In recent years, there has been a noticeable uptick in real estate prices, especially in the coastal regions, where demand for residential properties is high. In terms of property types, single-family homes are the most common, followed by condos, two-family homes, three-family homes, and four-family homes. It was found that the prices of multi-family homes have generally been lower than those of single-family homes, with some exceptions.

The analysis also revealed some significant regional variations in real estate sales patterns. The southwestern part of the state is home to some of the highest valued properties, while the eastern part of the state has generally lower property values. There are also some interesting differences in property types sold and sales volumes across different regions.

However, our study has several *limitations*. First, our analysis was based solely on sales data and did not take into account other factors that may influence the real estate market, such as demographic changes, economic conditions, or policy interventions. Second, our dataset did not include information on rental prices, which could provide additional insights into the health of the real estate market. Finally, our analysis focused exclusively on Connecticut and did not explore real estate sales patterns in neighboring states or regions.

Future research could address some of these limitations and extend our findings in several ways. For example, future studies could use a more comprehensive dataset that includes both sales and rental prices, as well as other relevant variables such as employment rates, household incomes, and population demographics. Such studies could also explore the impact of policy interventions on the real estate market, such as zoning regulations, tax incentives, or infrastructure investments. Finally, future research could compare real estate sales patterns across different states or regions to identify similarities and differences in market dynamics.

In *conclusion*, our study provides a comprehensive analysis of real estate sales patterns in Connecticut over several decades. Our findings suggest that real estate prices have fluctuated significantly over time, with noticeable regional variations and differences across property types. While our study has some limitations, it points to several important directions for future research, which could help policymakers and investors better understand the dynamics of the real estate market in Connecticut and beyond.

REFERENCES

[1] USA, "DATA.GOV: The Home of the U.S. Government's Open Data," US Government, [Online]. Available: https://data.gov/. [Accessed 5 May 2023].

[2] N. A. o. R. (NAR), "Highlights From the Profile of Home Buyers and Sellers," NAR, 2021. [Online]. Available: https://www.nar.realtor/research-and-statistics/research-reports/highlights-from-the-profile-of-home-buyers-and-sellers. [Accessed 7 May 2023].

[3] FRED, "Purchase only House price index for the United States," 2023 April 25. [Online]. Available: https://fred.stlouisfed.org/series/HPIPONM226S. [Accessed 7 May 2023].

[4] N. A. o. R. (NAR), "Existing-Home Sales," NAR, 2019. [Online]. Available: https://www.nar.realtor/research-and-statistics/housing-statistics/existing-home-sales. [Accessed 7 May 2023].

[5] Zillow, "Housing Data: Zillow Research," 2011. [Online]. Available: https://www.zillow.com/research/data/. [Accessed 7 May 2023].

[6] S. o. Connecticut, "Real Estate Sales 2001-2020 GL," data.ct.gov, [Online]. Available: https://catalog.data.gov/dataset/real-estate-sales-2001-2018. [Accessed 6 May 2023].