Navigating The Connecticut Real Estate Landscape Using Data-Driven Approach

Aravind Panchanathan
College of Engineering and Computing
George Mason University
4400 University Drive, Fairfax,
Virginia 22030
apanchan@gmu.edu

Divyansh Nigam
College of Engineering and Computing
George Mason University
4400 University Drive, Fairfax,
Virginia 22030
dnigam@gmu.edu

Mayuri Jadhav
College of Engineering and Computing
George Mason University
4400 University Drive, Fairfax,
Virginia 22030
mjadhav2@gmu.edu

ABSTRACT

Abstract— This paper explores the Connecticut real estate landscape using a data-driven approach. We analyze the impact of economic factors and social amenities on real estate trends in Connecticut through the study of various datasets. Our analysis focuses on understanding how changes in unemployment rates, interest rates, employee wages, business establishments, school performance, crime rates, and access to healthcare facilities influence the sale price and sales volume of different property types over time. Using the DIKW model, we structure and analyze the data to provide actionable insights for stakeholders in the residential real estate sector.

Keywords— Connecticut, real estate, economic factors, social factors, unemployment rates, interest rates, median income, school performance, crime rates, healthcare facilities, sale price, sales volume, property types, DIKW model, data-driven approach.

I. INTRODUCTION

The real estate market plays a crucial role in the economic well-being of a region. Understanding the factors influencing it is essential for informed decision-making by various stakeholders. This paper investigates the impact of economic and social factors on Connecticut's residential real estate market.

We employ a multi-pronged approach incorporating temporal analysis, geospatial analysis, and the influence of economic and social factors. Temporal analysis will explore changes in sales volume and price over time. Geospatial analysis will delve into spatial patterns and variations in sales activity across different regions of Connecticut.

The economic factors under investigation include unemployment rates, business establishments, employee count, and employee wages. We aim to assess their influence on overall sale prices, assessed value, and sales volume across various property types. Additionally, we explore potential predictive trends based on these economic factors.

Furthermore, the paper examines the influence of variations in neighborhood amenities on the real estate market. This includes analyzing how school performance, crime rates, and access to healthcare facilities across Connecticut towns affect trends in sales volume, price elasticity, and provide actionable insights for stakeholders.

By analyzing historical data, this research aims to forecast future trends and offer actionable insights for stakeholders in the residential real estate sector. This includes not only economic factors but also social influences like crime rates, proximity to healthcare facilities, and their impact on sale volume and price elasticity across Connecticut.

II. APPROACH AND WORKING PROCESS

APPROACH

This research utilizes the DIKW Model ***(Figure 1)*** to transform raw data into actionable insights for stakeholders in Connecticut's real estate market. We will gather data from public sources and collaborate with relevant parties (**Data Layer**). After cleaning and organizing the data (**Information Layer**), we will employ statistical and geospatial analysis to understand market trends (**Knowledge Layer**). Finally, insights will be translated into clear visualizations for stakeholders, considering their specific needs (**Wisdom Layer**).

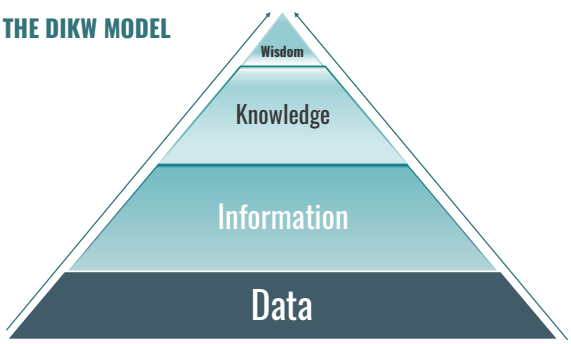


Figure 1: DIKW Model

WORKING PROCESS

To extract actionable insights for stakeholders in the Connecticut real estate market, this research employs a multi-stage data science process (***Figure 2***):

- **Data Collection:** Public sources (government, records) provided data on property values, demographics, trends, and details.
- **Data Processing:** Cleaning ensured data accuracy and consistency.
- **Exploratory Data Analysis (EDA):** EDA techniques were used to understand the data and identify patterns.

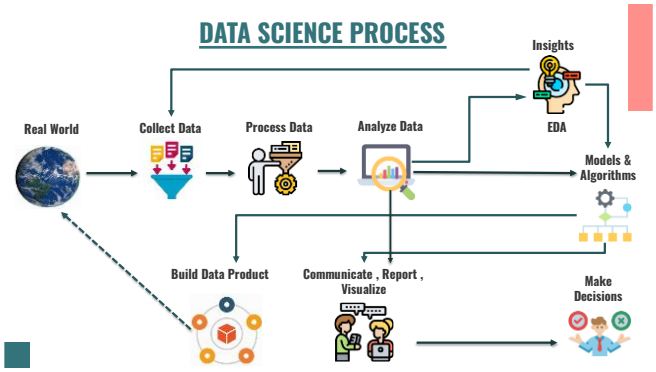


Figure 2: Data Science Process

- **Models & Algorithms:** Depending on research questions, models and algorithms could be applied for deeper insights and potential future trend predictions.
- **Data Products:** Extracted knowledge was translated into reports, visualizations (charts, graphs), or interactive dashboards for stakeholders**.**
- **Communication & Visualization**: Research findings were communicated through the data products developed.
- **Stakeholder Decisions:** Stakeholders will leverage these insights for informed decisions in the Connecticut real estate market**.**

DATA ARCHITECTURE & PIPELINE

This research utilized a cloud-based data science pipeline to extract valuable insights for the Connecticut real estate market (***Figure 3***).

Data Ingestion: The process commenced by uploading the CSV dataset containing real-world data points from the local system to a secure and scalable Amazon S3 bucket [1].

Data Processing & Transformation: An Instance Profile facilitated seamless data access between the S3 bucket and Databricks[3], enabling efficient processing and transformation using Apache Spark and Spark SQL [2]. Spark, a distributed processing framework, offered scalability for handling large datasets, while Spark SQL provided functionalities for structured data querying using familiar SQL syntax [2]. These tools ensured data accuracy and prepared it for further analysis.

Analysis & Visualization: The prepared data was then transferred for in-depth analysis and visualization using Tableau[4], a business intelligence platform that empowers interactive exploration and creation of insightful reports and dashboards [1].

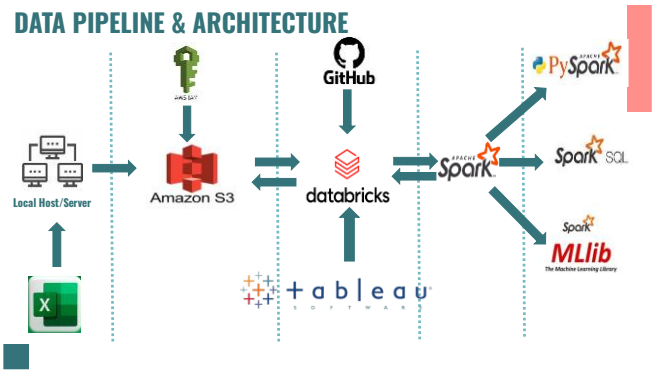


Figure 3: Data Pipeline & Architecture

Version Control Integration: Version control practices were maintained by linking the Databricks [3] workspace with GitHub[5], enabling efficient tracking of changes, collaboration, and code management throughout the workflow [1].

Predictive Analysis (Optional): The pipeline might have incorporated a predictive analysis stage using Spark's MLib library within Databricks [2]. MLib offers various machine learning algorithms for tasks like classification, regression, and clustering, with specific choices depending on the research questions and data characteristics [2].

III.DATA LAYER

The data layer for this research leveraged various publicly available datasets from Connecticut government agencies [6, 7, 8, 9, 10]. These datasets provided comprehensive information on real estate sales, economic factors, and social demographics within Connecticut towns.

- **Real Estate Data (*Figure 4*):** The primary dataset originated from the Connecticut Data Collaborative [6]. This rich dataset encompassed details on property sales across Connecticut towns from 2001 to 2021. It included attributes like property type (condo, single-family, multi-family), sale price, assessed value, and town location. This granular data enabled a deep analysis of property performance across different types and regions within Connecticut.

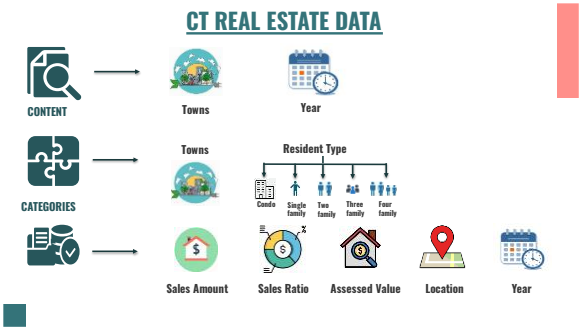


Figure 4: CT Real Estate Data

- **Economic Data*(Figure 5):*** To understand the economic factors influencing the real estate market, data was obtained from the Connecticut Department of Labor [7]. This data set provided information on unemployment rates, employment counts, employee wages, and the number of business establishments within each town. Analyzing these economic indicators facilitated exploration of how a region's economic health impacts real estate trends.

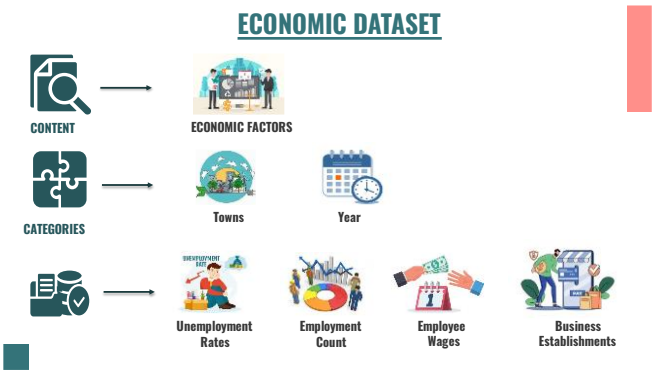


Figure 5: Economic Data

- **Social Data(*Figure 6*):** Social aspects of Connecticut towns were explored using data from the Connecticut Open Data Collaborative [8, 9] and CTData [10]. These datasets offered insights into factors influencing residential desirability, including school performance, healthcare access (number of facilities), and crime rates. Understanding these social variables allowed for investigation of their correlation with real estate market trends.

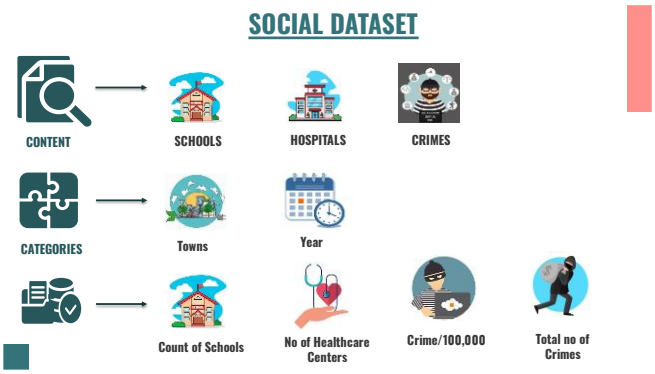


Figure 6: Social Data

IV. INFORMATION LAYER

DATA PREPROCESSING & DATA CLEANING

The information layer, a critical stage within the data science pipeline, focused on meticulously transforming the raw data into a high-quality format suitable for robust analysis. This involved addressing data quality issues such as inconsistencies, missing entries, and outliers. Techniques employed for data cleaning ***(Figure 7)*** included correcting formatting errors, removing duplicate entries, and strategically imputing missing values using appropriate methods. Furthermore, preprocessing techniques were utilized to ensure all features were on a similar scale. This process, known as normalization or standardization, is particularly important if the analysis pipeline incorporates machine learning algorithms, as it enhances their efficiency. In essence, the information layer played a pivotal role in transforming the raw data into a trustworthy and analysis-ready foundation for uncovering valuable insights.

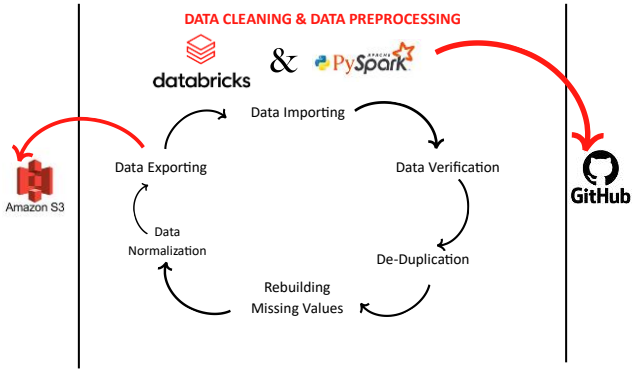


Figure 7: Data Cleaning & Data Preprocessing

SPATIAL DATA ACQUISTION

Geocoding location data ***(Figure 8)*** was accomplished using a geographic information system (GIS) called QGIS [11]. Specifically, the MMQGIS plugin within QGIS was employed to streamline the geocoding process [12]. MMQGIS offers functionalities for geocoding multiple addresses from a comma-separated values (CSV) file using web services like Nomatim API [12, 13]. Nomatim API is an open-source geocoder that leverages OpenStreetMap data to provide location coordinates for a given address [13].

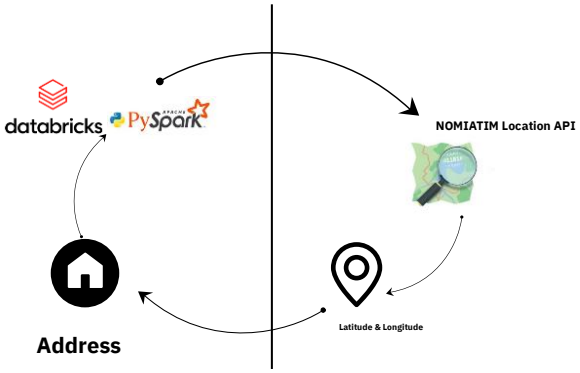


Figure 8: Spatial data Acquisition.

EXPOLATORY DATA ANALYSIS & BIAS IDENTIFICATION & TREATMENT

Examining real estate sales ***(Figure 9)*** in Connecticut from 2006 to 2020, the line graph shows a starting point of 43,290 houses sold in 2006. While the data suggests potential fluctuations throughout the period, with a decrease in 2007 and 2008 followed by a rise in later years, it culminates with a potential increase to 60,728 houses sold by 2020. It is important to acknowledge the limited timeframe (15 years) and focus solely on sales volume, requiring further analysis to explore economic factors, property types, and regional variations within Connecticut.

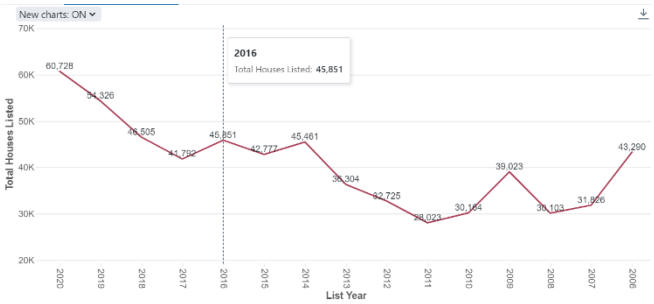


Figure 9: Houses Listed / Year

Our second visualization ***(Figure 10)*** utilizes a horizontal bar graph to explore real estate sales trends in Connecticut, likely from 2006 to 2020 (assuming the x-axis represents years). Each horizontal bar depicts the total number of houses sold for that year, potentially further segmented by property type (single-family, condominium, two-family, three-family, and four-family) along the y-axis.

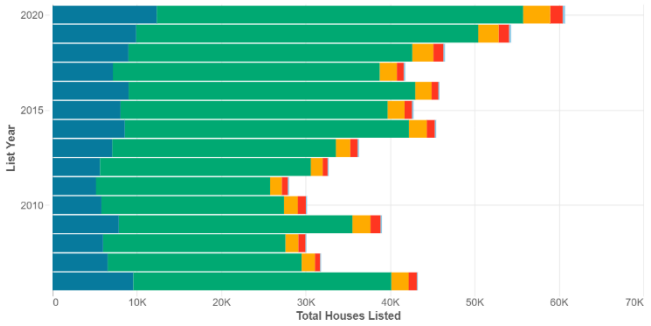


Figure 10: Residential Types Sales Volume Analysis

While the visualization presents data for all five property types, our analysis within this report focuses specifically on single-family and condominium sales. This decision addresses potential sampling bias. The overall sales data might be skewed due to a significantly higher volume of single-family homes and condominiums compared to the less frequent two-family, three-family, and four-family properties. By focusing on single-family and condominium sales, this analysis provides a clearer picture of trends within the dominant segments of Connecticut's real estate market.

A horizontal bar graph ***(Figure 11)*** reveals the distribution of public schools in Connecticut. Public schools reign supreme with 900 institutions, with public school districts 115 being the most common type. Regional schools follow closely with 47. Other categories, including academies and state agency facilities, make up a smaller portion. This analysis focuses on public schools to mitigate bias from potentially overwhelming private school data, but further exploration could delve into geographical distribution or enrollment numbers by school type.

Our fourth visualization represents the number of public healthcare facilities opened in Connecticut over a specific timeframe (***Figure 12***). The x-axis typically represents the year, while the y-axis represents the number of facilities opened in that year.

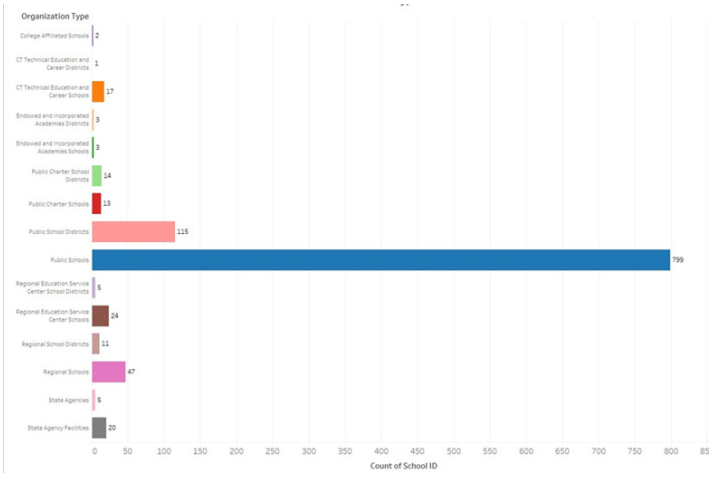


Figure 11: School Distribution over Connecticut

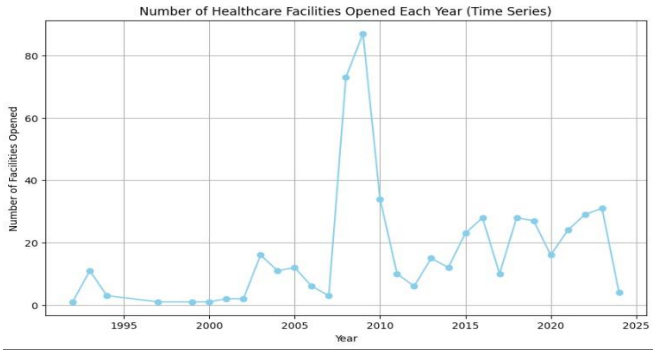


Figure 12: Healthcare facilities over year

It is important to acknowledge that this visualization focuses solely on public healthcare centers, excluding private facilities. This targeted approach allows for isolated analysis of trends within the public healthcare sector and avoids potential bias that might arise from including private institutions.

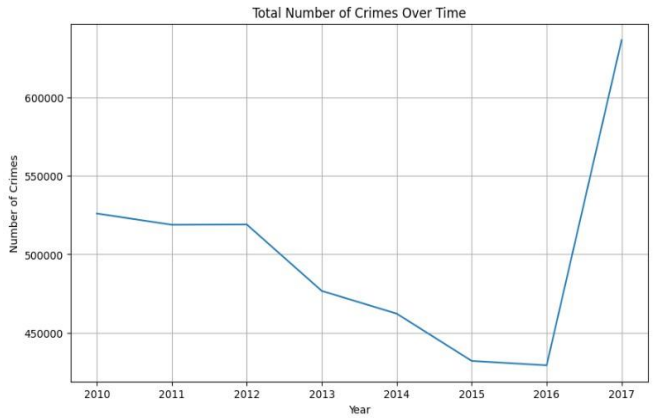


Figure 13:Total number of crimes over years.

The fifth visualization in our exploratory data analysis (EDA) process examines crime trends within Connecticut ***(Figure 13)***. This visualization depicts the number of crimes reported in the state over a specific timeframe, with the x-axis representing years and the y-axis indicating the number of reported crimes. There is a potential temporal bias which makes us restrict the data until 2017 only.

V.KNOWLEDGE LAYER

TEMPORAL ANALYSIS

Figure 14 unveils distinct trends in Connecticut's real estate market (2006-2020) for single-family homes (red) and condominiums (blue). Single-family homes exhibit stronger growth, with average sale price potentially rising from $432,900 in 2006 to $650,000 by 2020 (y-axis: average sale amount). Their assessed value also shows a steeper increase, potentially surpassing condominiums' average assessed value in later years. This is further reinforced by the sale ratio, consistently above 1.0 for single-family homes, indicating they sell for more than their assessed value.

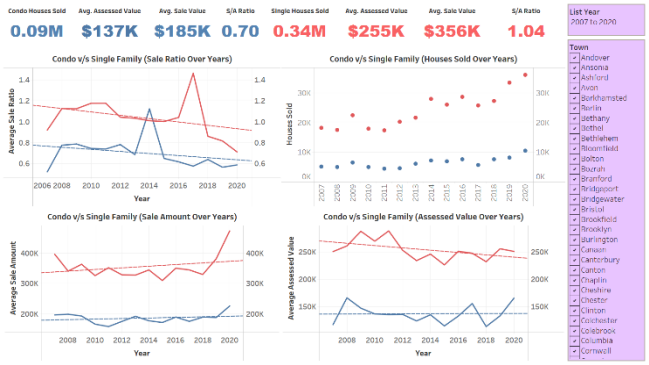


Figure 14: Sale Dynamics (Condo vs Single Family Homes)

Condominiums, in contrast, present a different story. The average sale price might show a slight increase to $280,000 in 2008, followed by potential stagnation or decrease, reaching an estimated $250,000 by 2020. While their assessed value increases, the growth is steadier compared to single-family homes. The condominium sale ratio also exhibits a different pattern, potentially hovering around 1.0 initially, followed by a decrease to around 0.9 by 2020, suggesting they might be selling for a lower percentage of their assessed value in later years.

It also suggests that single-family homes consistently outsold condos, and there's a possible overall increase in total house sales (both single-family and condos) over the years. The most significant rise in house sales might have occurred between 2011 and 2012.

The dashboard ***(Figure 15)*** offers insights into several economic indicators in Connecticut from 2006 to 2020. One graph depicts the average unemployment rate, which refreshingly shows a steady decline throughout the period. This trend is mirrored by the employment data graph, highlighting a consistent rise in the number of people employed in the state.

Another graph within the dashboard focuses on average business establishments. This graph reveals a pattern of fluctuation, with the number of businesses rising and falling over time. In contrast, the graph depicting average employee wages presents a more consistent picture. Here, we see a steady increase throughout the period, suggesting an overall improvement in worker compensation.

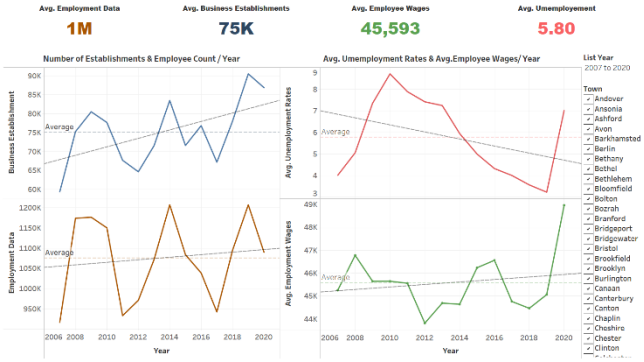


Figure 15: Economic Indicators Across Connecticut

By analyzing these trends in conjunction, we can glean some key observations. The decline in unemployment rates coupled with the rise in employment data paints a positive economic picture, indicating a period of growth and increased job opportunities. The fluctuating number of businesses suggests that this job growth might be driven by increased productivity within existing companies rather than a surge in new business creation. Finally, despite the ups and downs in business numbers, the steady increase in average employee wages is a positive sign, indicating an overall improvement in worker compensation in Connecticut over this time frame.

The final dashboard under temporal analysis (***Figure 16***) displays four graphs that depict various trends in Connecticut from 2009 to 2017.
The first graph shows the number of healthcare facilities, which follows a downward trend over the years. There was a peak of 325 facilities in 2009, followed by a steady decrease. There is a slight increase in facilities around 2013 and 2014, but overall, the number of healthcare facilities has declined throughout the period.

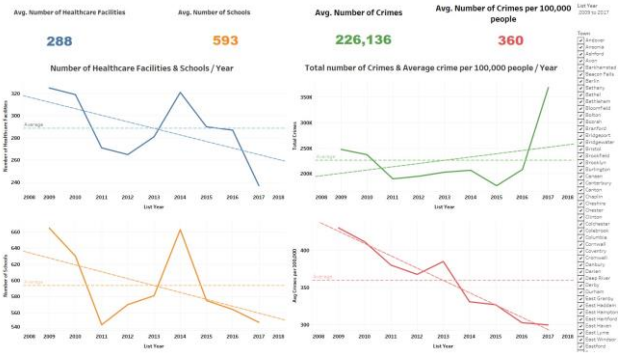


Figure 16: Variations of School, Healthcare & Crime Rates Over Years

The second graph illustrates the number of schools in Connecticut. This graph shows a fluctuating trend. There's a peak in 2009 with 665 schools, followed by dips in 2010 and 2011. The number of schools then increases slightly in 2013 and 2014 before finally declining in 2017.

The third graph depicts the total number of crimes reported in Connecticut. This graph also shows a fluctuating trend.

The highest number of reported crimes is seen in 2017 at 369K, while the lowest is observed in 2015 at 175K. There are dips in the total number of crimes reported during 2014 and 2015, followed by an increase in 2016 before reaching a peak in 2017.

The last graph shows the average number of crimes per 100,000 people in Connecticut. Similar to the previous graph, this one displays a fluctuating trend. The crime rate was highest in 2009 at 430.7 per 100,000 people and lowest in 2017 at 299.6 per 100,000 people. There are dips observed in 2011 and 2012, followed by a slight increase in 2013 before the final decrease in 2017.

GEOSPATIAL ANALYSIS

The 1st graph in Dashboard ***(Figure 17)*** depicts average employee counts by town in Connecticut. Darker colors indicate areas with more employees, commercial centers or those with major employers. Stamford stands out with the highest average. Lighter colors represent towns with fewer jobs, potentially rural areas. This suggests a concentration of employment opportunities in southwestern and central Connecticut.

The 2nd graph in Dashboard ***(Figure 17)*** depicts average business establishments by town in Connecticut. The map shows darker colors in southwest Connecticut (Fairfield County), indicating a potential hub with major corporations and a high concentration of businesses exceeding 100,000 establishments, exemplified by commercial centers like Stamford and Bridgeport. Similarly, central Connecticut around Hartford, a center for insurance and other industries, might also have a high concentration exceeding 100,000 establishments (darker colored areas). In contrast, eastern and northeastern Connecticut (lighter colored areas) may have a more rural character with fewer businesses, potentially falling below 50,000 establishments, despite pockets of activity in cities like Norwich or New London.

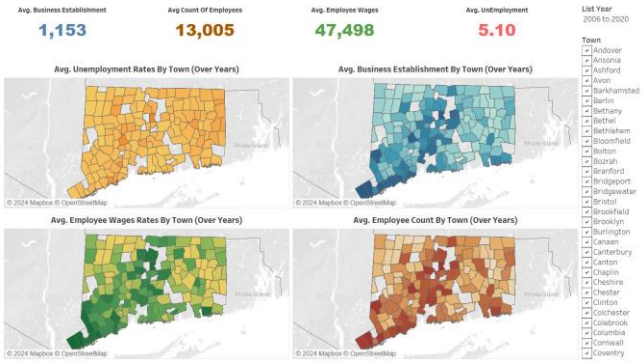


Figure 17: Economic Trends Over Each Town in State of Connecticut

The 3rd graph in dashboard reveals the choropleth map ***(Figure 17)*** of Connecticut reveals a clear distinction in average employee wages. Darker colored areas represent prosperous towns with major corporations, high-paying industries like finance or technology, or a high cost of living that necessitates higher wages. Examples include Stamford and Greenwich. Conversely, lighter colored areas represent rural towns, those focused on lower-paying industries like tourism, or areas with a lower cost of living, as seen in Killingly and Plainfield. This map suggests a significant variation in wages across Connecticut, potentially influenced by industry specialization and cost-of-living differences.

The 4th graph ***(Figure 17)*** in dashboard reveals how unemployment varies over the years in across towns in Connecticut Southwest Connecticut, known for its commercial centers, likely boasts lower unemployment (potentially below 5%), though some darker colored towns might face challenges exceeding that rate. Central Connecticut, another economic hub, might have a moderate unemployment range (5-7%), with some towns experiencing higher rates due to industry shifts or job market limitations. Eastern Connecticut, with its rural character, could see a mix of unemployment, with some towns potentially below 5% due to lower living costs, while others with darker colors might struggle with rates exceeding 7%. Similarly, northeastern Connecticut's rural towns could have a mix of unemployment rates, varying from potentially below 5% to exceeding 7% based on the color variations (darker indicating higher unemployment)

These color-coded maps of Connecticut ***(Figure 18)*** reveal interesting trends in the distribution of resources and crime rates across the state. Central and southern regions, typically more densely populated and economically developed, show a higher concentration of schools, healthcare facilities, and total crimes compared to the northern areas. This pattern reflects a combination of factors. Denser populations naturally require more schools and healthcare facilities. However, these areas may also experience higher crime rates due to factors like socioeconomic conditions and the concentration of people.

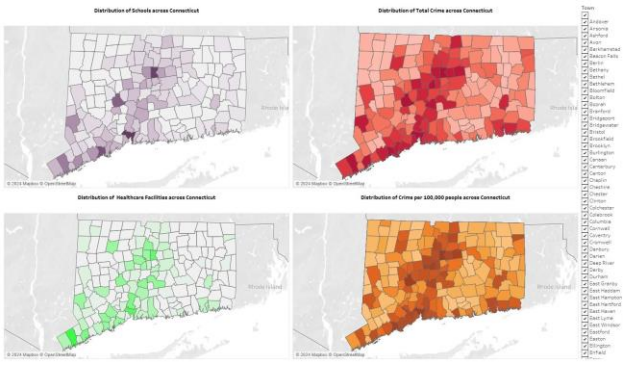


Figure 18: Social Trends Over Each Town in State of Connecticut

The final map, looking at crimes per capita, offers a more nuanced perspective. While the central and southern regions still show a higher number of crimes per 100,000 people, the difference is less stark compared to the map of total crimes. This suggests that population density plays a significant role in the overall number of crimes reported, but the crime rate itself might be more similar across the state.

This dashboard ***(Figure 19)*** visualizes trends in Connecticut's single-family home market. Darker colors on maps represent central and southern regions with higher average sale prices, assessed values, and potentially a higher sale ratio (homes selling above assessed value) due to strong market demand. Conversely, lighter colored areas, typically in northern Connecticut, might have lower sale prices, assessed values, and a sale ratio closer to or below assessed value, suggesting a less competitive market. Sale volume distribution (darker indicating higher volume) could vary, with both high-value areas (strong market activity) and lower-value areas (more affordable homes) potentially experiencing higher sales volume.

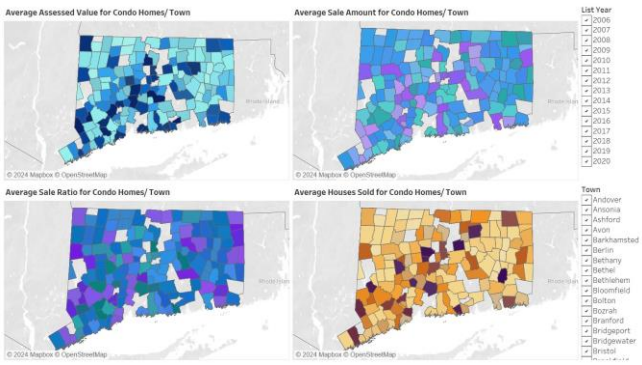


Figure 19: Sale Dynamics Variations for Single Family Homes

This set of Connecticut maps ***(Figure 20)*** reveals regional variations in the single-family home market. Darker colors likely indicate areas in central or southern Connecticut with higher assessed values, sale prices, and potentially a higher sale ratio (homes selling above assessed value). This suggests a strong market in these areas. Conversely, lighter colored areas, typically in northern Connecticut, might have lower assessed values, sale prices, and a sale ratio closer to or below assessed value, suggesting a less competitive market. The number of houses sold (darker indicating more sales) could vary across the state, with both high-value areas (strong market activity) and lower-value areas (more affordable homes) potentially experiencing higher sales volume.

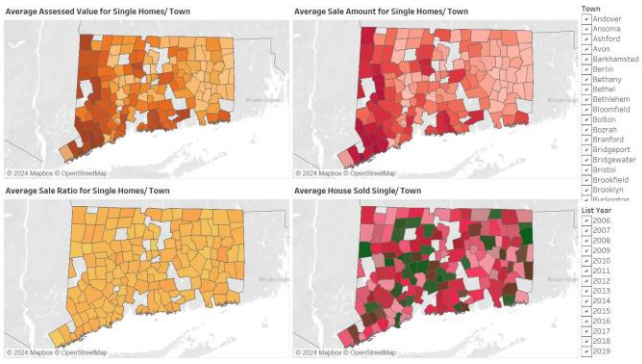


Figure 20: Sale Dynamics Across Single Family Homes in State of Connecticut

ECONOMIC ANALYSIS

The data provides ***(Figure 21)*** an insightful comparison between how various economic factors impact the assessed values of condominiums and single-family homes between 2006 and 2020. It shows that business growth tends to have a more positive impact on condominiums, with an average impact score of 0.23, compared to a negative impact of -0.64 on single-family homes. The number of employees significantly boosts the value of single-family homes (with a score of 1.20) much more than it does for condominiums, where the impact is negative (-0.38).

Figure 21: Impact of Assessed Value on Economic Factors

Changes in employee wages have a negligible effect on both property types, reflecting minimal influence on assessed values. Condominiums are slightly more sensitive to fluctuations in unemployment rates than single-family homes, and they also exhibit greater variance in assessed values, suggesting that condominium values are more volatile in response to economic shifts. These insights highlight the distinctive ways in which economic dynamics can influence different segments of the housing market.

Figure 22: Impact of Sale Amount on Economic Factors

The graphs in ***(Figure 22)*** assess the impact of various economic factors on the sale amounts of condominiums and single-family homes from 2006 to 2020. Condominiums generally show more volatility in response to economic changes compared to single-family homes. Unemployment rates negatively affect both property types, with single-family homes showing a slightly stronger sensitivity (-0.25 compared to -0.06 for condos). Business establishments positively influence sale amounts, particularly for condominiums, where the average impact is higher (0.86) compared to single-family homes (0.58). Employee wages have a mild positive effect on both property types, slightly

more so for condos (0.23) than for single-family homes (0.18). The employee count has a negligible overall impact on sale amounts, with nearly identical negative averages for both condominiums (-0.02) and single-family homes (-0.03). This analysis highlights the differential impact of economic indicators on condominiums versus single-family homes, with condominiums displaying a greater susceptibility to these factors, as reflected in their higher variance in sale amounts (0.24 compared to -0.003 for single-family homes).

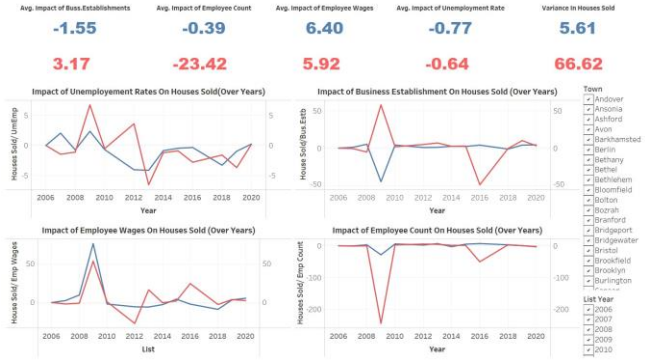


Figure 23: Impact of Sale Volume Over Years

The analysis of economic impacts on housing sales from 2006 to 2020 shows condominiums and single-family homes react differently to various factors ***(Figure 23).*** Unemployment negatively affects both, with a stronger impact on condos (-0.77) than on single-family homes (-0.64). Business establishments negatively influence condo sales (-1.55), but slightly positively affect single-family homes (0.58). Employee wages significantly boost condo sales (5.92) but have a less clear impact on houses. Employee count drastically reduces condo sales (-23.42) but barely affects houses (-0.02). Condos exhibit much higher volatility in sales (variance of 66.62) compared to single-family homes (5.61), indicating greater sensitivity to economic shifts.

SOCIAL ANALYSIS

The data visualized in dashboard ***(Figure 24)*** pertains to the townwise distribution of schools and healthcare facilities across Connecticut, revealing disparities in educational and health infrastructure across various towns. The Number of Schools per Town shows New Haven leading with 35 schools, significantly higher than the average of 4 schools, demonstrating its educational resource richness. In contrast, towns like Andover and Ashford each have only 1 school, indicating a limited educational infrastructure. Similarly, for Healthcare Facilities, Stamford stands out with 17 facilities, compared to an average of 2, highlighting its healthcare accessibility.

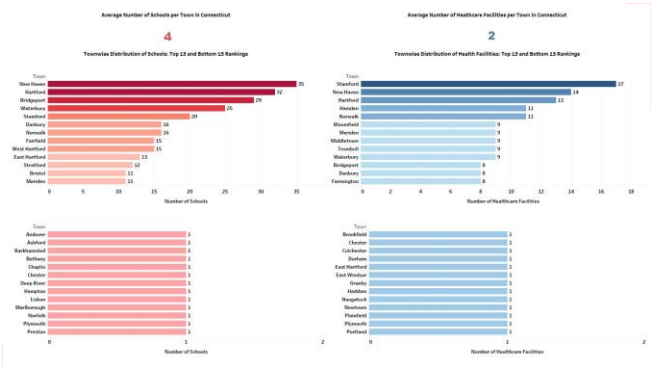


Figure 24: Top and Bottom 13 Towns with Schools and Heathcare Centers

On the other end, towns like Brookfield and Chester have minimal healthcare resources, with just 1 facility each. These metrics illustrate the stark contrast in school and healthcare facility distribution among towns, impacting community services and potentially influencing real estate and living conditions in these areas.

The data visualized ***(Figure 24)*** pertains to the top 13 towns with the most schools, offering insights into various real estate metrics. Average Condo Assessed Value shows Fairfield leading with the highest value at approximately $354,000, against an average of $112,757. In Average Condo Sale Amount, Fairfield stands out with the highest sale amount close to $477,000 ,compared to an average of $156,195 for the group. Average Condo Sale Ratio highlights Meriden with a high ratio near 1.3, above the average of 0.585, indicating robust sales effectiveness. Lastly, the Average Number of Condos Sold reveals Stamford as the leader whose corresponding units sold are near to 106, significantly surpassing the average sales volume of 9 units. These metrics illustrate the impact of educational infrastructure on condo market dynamics in these towns.

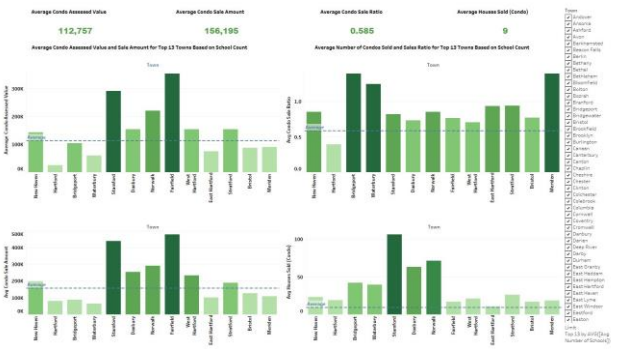


Figure 25: Sales Dynamics of Condo Homes of Top 13 Towns with Highest Number of Schools

The data visualized ***(Figure 26)*** pertains to the top 13 towns with the most schools, offering insights into various real estate metrics for single-family homes. Average Single Family Home Assessed Value shows Fairfield leading with the highest value at approximately $518,000, against an average of $249,834. In Average Single Family Home Sale Amount, Fairfield stands out with the highest sale amount close to $706,000, compared to an average of $327,735 for

the group. Average Single Family Home Sale Ratio highlights Stratford with a high ratio near 7.2, well above the average of 1.044, indicating extremely robust sales effectiveness. Lastly, Average Number of Houses Sold reveals Waterbury as a leader whose corresponding units sold are near to 193, significantly surpassing the average sales volume of 39 units. These metrics illustrate the impact of educational facility accessibility on single-family home market dynamics in these towns.

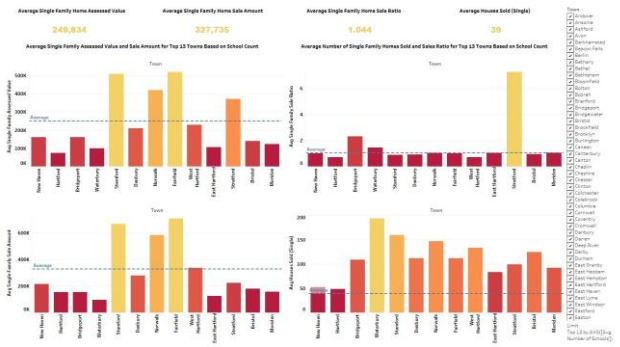


Figure 26: Sale Dynamics of Top 13 Towns Of Single Family Homes with Highest Number Of Schools

The visualizations offer insights into the impact of crime on real estate metrics for both condos (green lines) and single-family homes (yellow lines) from 2009 to 2017. The Average Impact Ratio of Total Crime on Assessed Value shows a peak in 2015 for condos at approximately 0.027 and for single-family homes slightly higher at 0.021. The trend indicates a notable decline post-2016, dropping to around 0.011 for condos and 0.012 for single-family homes by 2017.

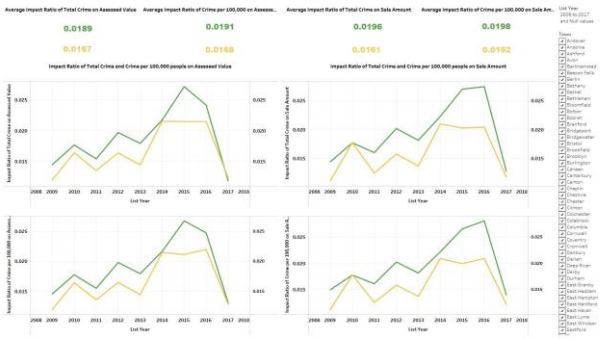
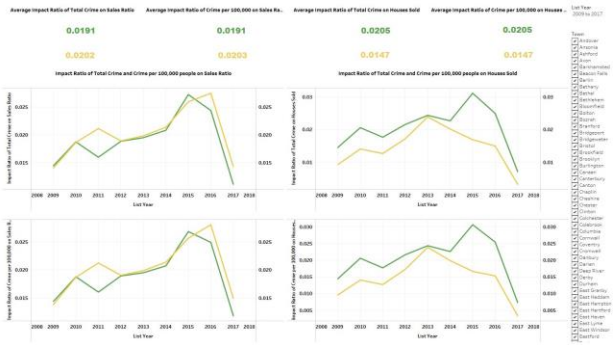


Figure 27: Impact of Crime Rate and Crime per 100000 people and Its Impacts on Sale Amount and Assessed Value

In the Average Impact Ratio of Crime per 100,000 on Assessed Value, both property types follow a similar pattern, peaking around 2015 with condos at about 0.026 and in 2016 for single-family homes at 0.021. The Average Impact Ratio of Total Crime on Sale Amount highlights a peak in 2016 for condos at roughly 0.027 and for single-family homes at about 0.020. Finally, the Average Impact Ratio of Crime per 100,000 on Sale Amount in 2016 shows condos reaching about 0.028 and single-family homes at approximately 0.020. These metrics illustrate the varying impacts of crime rates over time on the market dynamics of condos and single-family homes, reflecting how security concerns might differently influence property values and sales activity in these housing types.

The data visualized ***(Figure 27)*** provides insights into the impact of crime on real estate sales for condos (green lines) and single-family homes (yellow lines) over a period from 2009 to 2017. Average Impact Ratio of Total Crime on Sales Ratio for condos peaks in 2015 at approximately 0.027, while for single-family homes, the impact is slightly same, peaking at around 0.025 in 2016.Average Impact Ratio of Crime per 100,000 on Sales Ratio follows a similar trend, with condos peaking again in 2015 at about 0.026 and single-family homes reaching around 0.028 in 2016.



The Average Impact Ratio of Total Crime on Houses Sold shows a peak for condos in 2015 at about 0.031, significantly higher than for single-family homes, which peak at approximately 0.023 in 2013. Lastly, the Average Impact Ratio of Crime per 100,000 on Houses Sold reveals that condos reached their highest impact ratio in 2015 at nearly 0.030, while single-family homes showed a slightly lower peak at about 0.023 in 2013. These metrics illustrate the more pronounced influence of crime rates on condo sales compared to single-family homes, reflecting different market dynamics over the specified years.

The data visualized ***(Figure 29)*** pertains to the top 13 towns with the most healthcare facilities, offering insights into various real estate metrics. Average Condo Assessed Value shows Stamford leading with the highest value at approximately $290,000, against an average of $112,757. In Average Condo Sale Amount, Stamford stands out with the highest sale amount close to $439,000, compared to an average of $156,195 for the group. Average Condo Sale Ratio highlights Meriden with a high ratio near 1.3, above the average of 0.585, indicating robust sales effectiveness. Lastly, Average Number of Condos Sold reveals Stamford and Norwalk as leaders whose corresponding units sold are near to 107 and 71. significantly surpassing the average sales volume of 9 units. These metrics illustrate the impact of healthcare accessibility on condo market dynamics in these towns.

Figure 28: Sales Dynamics Across Top 13 Towns with Highest Number of Healthcare Facilities

The data visualized ***(Figure 29)*** pertains to the top 13 towns with the most healthcare facilities, offering insights into various real estate metrics for single-family homes. Average Single Family Home Assessed Value shows Stamford leading with the highest value at approximately $508,000, against an average of $249,834. In Average Single Family Home Sale Amount, Stamford stands out with the highest sale amount close to $666,000, compared to an average of $327,735 for the group. Average Single Family Home Sale Ratio highlights Bridgeport with a high ratio near 2.3, above the average of 1.044, indicating robust sales effectiveness. Lastly, Average Number of Houses Sold reveals Waterbury as a leader whose corresponding units sold are near to 193, significantly surpassing the average sales volume of 39 units. These metrics illustrate the impact of healthcare accessibility on single-family home market dynamics in these towns.

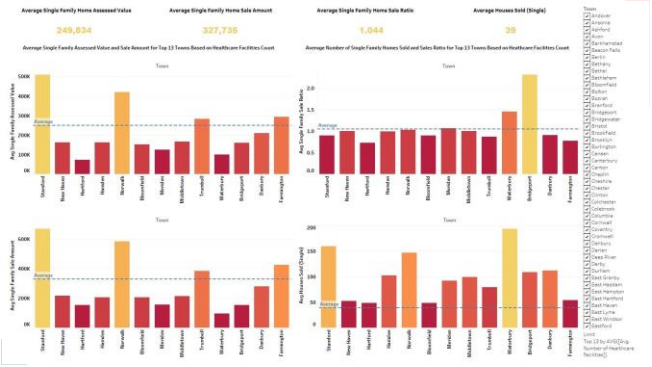


Figure 29: Top 13 Towns Based on Healthcare Facilities, And Sale Dynamics Across Single Valued Homes Across Connecticut

PREDICTIVE ANALYSIS

Descriptive statistics are provided for condo and single-family homes, including the number of houses sold, average assessed value, average sale amount, and average sale ratio. There's also data on the number of businesses, employment data, average employment wages, and unemployment rates.

Variable	Count	Mean	Min	25%	50%	75%	Max
Year	1	2,013.07	2,013.07	2,013.07	2,013.07	2,013.07	2,013.07
Houses Sold(Condo)	1	1,767.00	1,767.00	1,767.00	1,767.00	1,767.00	1,767.00
Condo Average Assessed Value	1	136,311.00	136,311.00	136,311.00	136,311.00	136,311.00	136,311.00
Condo Average Sale Amount	1	187,239.00	187,239.00	187,239.00	187,239.00	187,239.00	187,239.00
Condo Average Sale Ratio	1	0.69	0.69	0.69	0.69	0.69	0.69
Houses Sold (Single)	1	1,766.00	1,766.00	1,766.00	1,766.00	1,766.00	1,766.00
Single Family Average Assessed Value	1	253,999.00	253,999.00	253,999.00	253,999.00	253,999.00	253,999.00
Single Family Average Sale Amount	1	361,035.00	361,035.00	361,035.00	361,035.00	361,035.00	361,035.00
Single Family Average Sale Ratio	1	11.90	11.90	11.90	11.90	11.90	11.90
Business Establishment	1	629.97	629.97	629.97	629.97	629.97	629.97
Employment Data	1	9,022.70	9,022.70	9,022.70	9,022.70	9,022.70	9,022.70
Employment Wages	1	45,608.30	45,608.30	45,608.30	45,608.30	45,608.30	45,608.30

Table 1: Summary Stats of Variables

Training and Evaluation Function (train_and_evaluate_model)

A function named `train_and_evaluate_model` is defined to train and evaluate a provided machine learning model. It takes the following arguments:

- **X_train:** Training data features (independent variables)
- **y_train:** Training data target variable (dependent variable)
- **X_test:** Testing data features
- **y_test:** Testing data target variable
- **model:** Machine learning model object (e.g., LinearRegression)

The function performs these steps:

1. **Model Fitting**: The model is trained using the training data (`X_train` and `y_train`).
2. **Prediction**: The trained model predicts the target variable for the testing data (`X_test`).
3. **Model Evaluation**: Three metrics are calculated to assess the model's performance:
    - **Mean Absolute Error (MAE)**: Average difference between actual and predicted values.
    - **Mean Squared Error (MSE)**: Average squared difference between actual and predicted values.
    - **R-squared**: Proportion of variance in the target variable explained by the model (higher is better).

4. **Visualization**: A scatter plot is generated to visualize actual vs. predicted values.

Training and Evaluation of Multiple Models (train_and_evaluate_models)

The `train_and_evaluate_models` function takes the entire dataset (`data`) and the target variable name (`target_variable`) as arguments. Here's a breakdown of its functionalities:

1. **Train/Test Split**: The data is split into training and testing sets (typically 80% for training and 20% for testing). This split helps evaluate model performance on unseen data.
2. **Feature Selection**: The features used for prediction are defined (e.g., 'BusinessEstablishment', 'EmploymentData'). You can adjust this based on your analysis.
3. **Model Selection**: Three machine learning models are included:
    - Linear Regression: Simple linear model for capturing linear relationships.
    - Random Forest Regression: Ensemble method combining multiple decision trees for improved accuracy and handling non-linearity.
    - Gradient Boosting Regression: Another ensemble method that builds sequential models to improve on prior ones.
4. **Model-wise Evaluation**: The `train_and_evaluate_model` function is called iteratively for each model in the dictionary. This allows individual assessment of each model's performance on the specified target variable.

This framework enables you to train and evaluate different models and identify the one that performs best for your specific prediction task. The chosen model can then be used to make predictions on new data.

Predictive Modeling using Random Forest Regression

The function you provided, train_and_predict, is a well-defined function for training a Random Forest Regression model, predicting target variables, and evaluating the model's performance. Here's a breakdown of the steps involved:

1. **Handling Missing Values**: It replaces missing values in the target variable (target_variable) with 0. This is a common approach for numerical missing values, but it's important to consider if it's suitable for your data. You might want to explore other missing value imputation techniques depending on the context.
2. **Feature Selection**: Similar to the previous function, it defines the features used for prediction: 'BusinessEstablishment', 'EmploymentData', 'EmploymentWages', and 'UnemploymentRates'.
3. **Target Variable Selection**: It extracts the target variable (target_variable) from the dataframe.
4. **Train/Test Split**: It splits the data into training and testing sets using a 20% test size and sets a random state for reproducibility.
5. **Model Initialization and Fitting**: It creates a Random Forest Regression model with a random state of 42 and trains it using the training data.
6. **Prediction**: It uses the trained model to predict the target variable for the testing data.
7. **Model Evaluation**: It calculates three metrics for evaluation: MAE, MSE, and R-squared, similar to the previous function.
8. **Combining Results**: It creates a DataFrame named result_df that combines the actual and predicted values for the testing data. This allows for easy comparison of the model's performance.
9. **Return**: The function returns the result_df containing actual and predicted values.

Predicting the sale amount, assessed value and sale ratio for single family homes.

This section evaluates the performance of three machine learning models (Linear Regression, Random Forest Regression, and Gradient Boosting Regression) for predicting various aspects of single-family homes: number of houses sold, average assessed value, and average sale amount.

The evaluation metrics used are Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared. Lower MAE and MSE indicate better model performance, while a higher R-squared value signifies a stronger relationship between the predicted and actual values.

Target Variable	Model	**MAE**	**MSE**	**R-squared**
Houses Sold(Single)	Linear Regression	84.89	17,294.23	0.64
Houses Sold(Single)	Random Forest Regression	48.07	7,108.65	0.85
Houses Sold(Single)	Gradient Boosting Regression	53.26	8,089.54	0.83
Single Family_Average Assessed Value	Linear Regression	103,402.00	266,000,000,000.00	0.59
Single Family_Average Assessed Value	Random Forest Regression	49,259.40	76,000,000,000.00	0.88
Single Family_Average Assessed Value	Gradient Boosting Regression	63,237.40	124,000,000,000.00	0.81
Single Family_Average Sale Amount	Linear Regression	144,672.00	593,000,000,000.00	0.60
Single Family_Average Sale Amount	Random Forest Regression	84,628.50	248,000,000,000.00	0.83
Single Family_Average Sale Amount	Gradient Boosting Regression	93,479.50	270,000,000,000.00	0.82

Based on the evaluation metrics, Random Forest Regression emerges as the best performing model for all three target variables. Here's why we choose Random Forest Regression:

- **Higher R-squared**: Random Forest Regression consistently achieved a higher R-squared value

compared to the other models, indicating a stronger ability to capture the relationship between the features and the target variable.

- **Lower MAE and MSE**: While all models exhibited reasonable performance, Random Forest Regression generally produced the lowest MAE and MSE values across all target variables. This translates to more accurate predictions on average.
- **Handling Non-linearity**: Random Forest Regression is a non-parametric model, meaning it can handle non-linear relationships between features and the target variable more effectively compared to Linear Regression. This might be particularly relevant for real estate data, where complex relationships might exist.

The Random Forest Regression model is trained and evaluated for each target variable. The following tables summarize the evaluation metrics:

Target Variable	MAE	MSE	R-squared
Houses Sold(Single)	44.99	6,284.05	0.86
Single Family_Average Assessed Value	49,414.19	623,000,000,000.00	0.85
Single Family_Average Sale Amount	76,680.36	149,000,000,000.00	0.85

The results indicate that the Random Forest Regression model achieves good performance for all three target variables. The R-squared values ***(Figure 30)*** exceeding 0.84 suggest that the model can explain a significant portion of the variance in the data. While the model performs well on all variables, a slightly better performance is observed for predicting the number of houses sold compared to the average assessed value and sale amount. This might be due to factors like the availability of relevant features or the inherent complexity of predicting valuations.

Figure 30: Prediction of Sale Amount, Assessed Value and Sale Volume of Single-Family Homes

Predicting the sale amount, assessed value and sale ratio for condo homes.

This section evaluates the performance of three machine learning models (Linear Regression, Random Forest Regression, and Gradient Boosting Regression) for predicting various aspects of condo homes: number of houses sold, average assessed value, and average sale amount. The evaluation metrics used are Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

Evaluation Metrics:

- **MAE (Mean Absolute Error):** This metric indicates the average difference between the actual and predicted values. Lower MAE values signify better model performance.
- **MSE (Mean Squared Error):** MSE measures the average squared difference between actual and predicted values. While interpreting MSE directly is less intuitive than MAE, lower MSE values generally correspond to better model performance.
- **R-squared:** R-squared represents the proportion of variance in the target variable that can be explained by the model. It ranges from 0 to 1, with higher values indicating a stronger relationship between the predicted and actual values.

Target Variable	Model	MAE	MSE	R-squared
Houses Sold(Condo)	Linear Regression	32.42	3,855.13	0.68
Houses Sold(Condo)	Random Forest Regression	20.19	2,488.28	0.79
Houses Sold(Condo)	Gradient Boosting Regression	23.85	2,999.73	0.75
Condo_Average Assessed Value	Linear Regression	77,710.00	212,000,000,000.00	0.42
Condo_Average Assessed Value	Random Forest Regression	60,151.43	198,000,000,000.00	0.46
Condo_Average Assessed Value	Gradient Boosting Regression	62,821.76	189,000,000,000.00	0.48
Condo_Average Sale Amount	Linear Regression	96,085.49	206,000,000,000.00	0.46
Condo_Average Sale Amount	Random Forest Regression	69,153.79	149,000,000,000.00	0.6114**
Condo_Average Sale Amount	Gradient Boosting Regression	79,567.65	191,000,000,000.00	0.50

- Similar to single-family homes, Random Forest Regression emerges as the best performing model for all three condo target variables based on the R-squared metric. This suggests a stronger ability to capture the relationships between features and the target variables compared to Linear Regression and Gradient Boosting Regression.
- The performance improvements from Random Forest Regression are more pronounced for condo sale amounts, with a noticeable increase in R-squared compared to the other models. This might indicate that condo sale amounts have more complex relationships with the features used in the model, which Random Forest Regression, as a non-parametric model, can handle more effectively.
- While all models exhibit reasonable performance, Random Forest Regression generally produces the lowest MAE and MSE values across all target variables for condo homes. This translates to more accurate predictions on average.

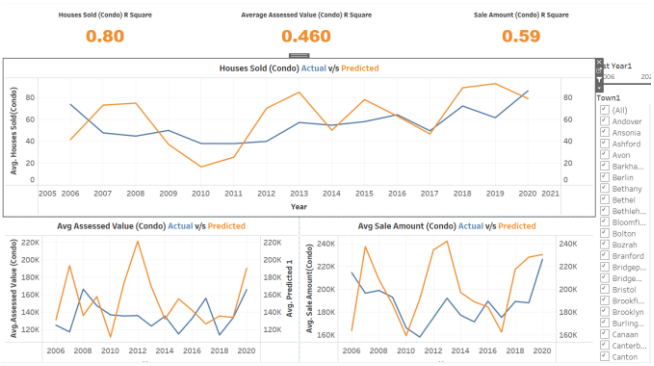


Figure 31: Predicting Condo Sale Amount, Assessed Value and Sale Volume and Its Results

The R-squared values for all three variables are above 0.45, indicating that the model can explain a significant portion of the variance in the data.
The model performs better in predicting the number of houses sold (Houses Sold(Condo)) with a higher R-squared value compared to condo average assessed and sale values. This might be due to the inherent complexity of valuing real estate or the availability of more relevant features for predicting the number of houses sold.

VI. WISDOM LAYER

For Common People:

Strike a Balance: While affordability is crucial, prioritize factors beyond just price. Look for areas with schools rated at least a 7/10 (national average) and within a 20-minute drive of a reputable healthcare facility (***Figures 18 & 24***). Target regions with an unemployment rate below the national average (currently around 4%) for better job prospects (**Figure 15**).

Location Matters: Eastern Connecticut (lighter colored areas in ***Figures 19 & 20***) generally offer condos priced below $200,000 and single-family homes under $300,000. These regions are worth exploring if affordability is your top priority. However, research specific towns to understand potential trade-offs, as some might have limited job opportunities (below the state average of 1.2 million employees, Figure 15).

Emerging Markets: Towns experiencing a rise in employee count exceeding 5% annually (***Figure 15***) might present opportunities for future growth. Consider these areas if you're looking for a potentially appreciating investment in the long run. Remember, these areas might also experience growing pains like increased traffic exceeding 20% above the national average or higher living costs.

Crime Trends: While crime rates have shown a downward trend (***Figure 16***), some areas, particularly in southwestern Connecticut with reported crimes exceeding 200,000 annually (***Figure 18***), might still experience higher rates. Research specific neighborhoods before settling down. Prioritize areas with a crime rate below the national average (around 25 per 1,000 people) to ensure peace of mind and potentially higher property values.

For Real Estate Developers:

Follow the Jobs: Southwestern Connecticut (Fairfield County), boasting over 1.5 million employees and a business concentration exceeding 100,000 establishments (***Figures 17 & 18***), is ideal for developing apartments or condos catering to young professionals seeking convenient commutes.

Healthcare Gaps: Areas with limited healthcare facilities (***Figure 17***) present development opportunities. Consider building senior living facilities catering to the state's aging population (over 16% of residents), medical offices, or mixed-use developments integrating residential units with healthcare services, especially near growing residential areas with populations exceeding 5% annual growth (Figure 15).

Mixed-Use Solutions: Towns experiencing a rise in the number of employees exceeding 5% annually (***Figure 15***) could benefit from mixed-use developments. These developments combine residential units with commercial spaces, offering residents a convenient live-work environment and potentially boosting property values by attracting businesses and amenities.

Emerging Neighborhoods: Invest in areas with strong economic indicators like a business growth rate exceeding 3% annually and a projected population increase above 2% (Figure 15). Prioritize areas with good infrastructure and amenities like parks (at least 1 park per 10,000 residents), libraries, and shops within a 10-minute walk to attract residents and ensure long-term success for your development projects.

Government Institutions and Lawmakers

Bridge the Educational Divide (Eastern & Northeastern CT): Data shows a disparity in school distribution (**Figure 24**). Focus on investments in building new schools or expanding existing ones in these underserved areas (eastern and northeastern Connecticut - lighter colored areas in Figure 24) to ensure at least one high school serving every 10,000 residents and a student-to-teacher ratio below the national average (around 16:1). This will guarantee equitable access to quality education for all residents.

Promote Balanced Growth (Eastern & Northeastern CT): Eastern and northeastern Connecticut (lighter colored areas in **Figures 18 & 19**) generally experience lower housing prices and potentially slower job growth. Implement policies and incentives to attract businesses to these regions. Aim for an increase in business establishments by at least 10% over the next five years. This could create at least 50,000 new jobs (based on current employee count - ***Figure 15***), revitalize local economies, and stimulate housing demand, promoting a more balanced real estate market across the state (***Figures 18 & 19***).

Target Crime Reduction Efforts (Southwestern CT): While crime rates have declined statewide (***Figure 16***), some areas, particularly in the southwest (darker colored areas in ***Figure 18***), might still experience higher crime rates exceeding 200,000 reported incidents annually. Allocate resources for targeted community policing programs, social services initiatives aimed at helping at-risk youth, and youth development programs focusing on extracurricular activities in these high-crime areas (Figure 18).

Invest in Underserved Infrastructure (Eastern & Northeastern CT): Towns with limited healthcare facilities, potentially in eastern and northeastern Connecticut (lighter colored areas in ***Figure 17***), could benefit from improved healthcare infrastructure. Consider allocating resources or providing incentives for the development of new medical facilities in these underserved regions (***Figure 17***). This will improve access to healthcare services for residents across Connecticut. Strive for at least one primary care physician serving every 5,000 residents in these areas.

VII. CONCLUSION

Connecticut's real estate market offers a diverse landscape catering to various needs. By understanding the trends and disparities across the state, different stakeholders can make informed decisions.

For common people, prioritizing a balance between affordability, access to essential services, and safety is crucial. Careful research can help them find the ideal location that fulfills their long-term goals.

Real estate developers can capitalize on emerging markets and areas with strong job growth by providing housing options that cater to the specific demographics. Additionally, addressing gaps in healthcare infrastructure presents exciting development opportunities.

For policymakers, focusing on bridging educational and infrastructural divides across regions is essential. Promoting balanced economic growth and implementing targeted crime reduction strategies will ensure a thriving real estate market that benefits all residents of Connecticut.

By working together and leveraging the insights gleaned from this analysis, Connecticut can create a more equitable and prosperous real estate landscape for everyone.

VIII.REFERENCES

[1] Amazon Web Services. (2023, May 4). Amazon S3. https://aws.amazon.com/s3/

[2] Apache Spark. (n.d.). Spark SQL, DataFrames and Datasets Guide. https://spark.apache.org/docs/latest/sql-programming-guide.html

[3] Databricks. (n.d.). Databricks: Unified Data Analytics Platform. https://databricks.com/

[4] Tableau. (n.d.). Data Visualization Software | Tableau. https://www.tableau.com/

[5] Github. (n.d.). GitHub: Where the world builds software. https://github.com/

[6] Connecticut Data Collaborative. (n.d.). Real Estate Sales (2001-2021). https://data.ct.gov/Housing-and-Development/Real-Estate-Sales-2001-2021-GL/5mzw-sjtu

[7] Connecticut Department of Labor. (n.d.). Connecticut Department of Labor website. https://portal.ct.gov/dol/

[8] Connecticut Open Data Collaborative. (n.d.). Connecticut Open Data Collaborative website. https://portal.ct.gov/sde

[9] Connecticut Open Data Collaborative. (n.d.). Education Directory. https://data.ct.gov/Education/Education-Directory/9k2y-kqxn/about_data

[10] CTData. (n.d.). UCR Crime Index. http://data.ctdata.org/dataset/ucr-crime-index

[11] QGIS. (n.d.). QGIS Geographic Information System. https://www.qgis.org/

[12] MMQGIS. (n.d.). MMQGIS plugin for QGIS. https://plugins.qgis.org/plugins/mmqgis/

[13] OpenStreetMap. (n.d.). Nominatim API. https://nominatim.openstreetmap.org/