Ryan Malone
CSCE567-001

April 21, 2025
Final Project Writeup

# Connecticut Real Estate Visualization Project

## Midterm Proposal Outline Update:

**Original Proposal:** The original midterm proposal that I submitted has been included in the zip file provided as "midterm_proposal.pdf"

**Instructor Feedback:** Nice work on the project proposal! Following your stated plan should result in an interesting and informative project. If any major changes occur, please let me know!

**Peer Feedback:** The feedback I got from the peer reviews were both very positive and didn't mention many things to change except for explaining why Connecticut was chosen in the overview. I've attached the peer feedback as "Peer_Review1.pdf" and "Peer_Review2.pdf" in the zip file provided to reference.

**Changes Made:**

After receiving the feedback that was mostly good from both the instructor and my peers, I decided to go ahead and get started on the project following my original proposal – I did not immediately implement any changes to it. However, after starting the project and working with the data in Tableau I did realize that some things need to change from my original proposal. For example, when working on the geographic map of Connecticut showing the number of sales within each town, I ran into issues with how Tableau was displaying the towns. It wasn't matching the layout or appearance I had originally planned for (as shown in the example images in my proposal). To fix this, I downloaded a separate dataset containing zip code data from https://www.unitedstateszipcodes.org/zip-code-database/, matched the zip codes to the towns in Connecticut, and used that to build the map instead. This allowed me to still show the number of properties sold by town, just using a slightly different approach than originally planned. The geographic map now shows each town, and the number of properties sold, however when hovering, you users can still hover over individual zip codes (the information remains the same though).

Another adjustment I had to make came when I realized that real estate serial numbers in the dataset were not actually unique (some sales had the same serial number for different properties/sales). Originally, I planned to show each sale individually on my scatter plot by using the serial number as a detail. However, I noticed that some sale prices appeared extremely inflated because Tableau was summing them when duplicates were present. To fix this, I combined the serial number with the town name in the detail section. This made it so each sale appeared only once and was accurately displayed in my visualization. I also originally hoped to include a toggle to turn the shading of economic downturns on and off, but since I had to use a drawn area on one of the graphs (and a reference band on another), I couldn't get the toggle working the way I wanted in Tableau.

Overall, I followed my project proposal very closely and didn't make many changes to the actual plan. Instead, I moved forward with my original proposal and when I ran into issues while building my visualizations, I adjusted by using calculated fields, alternate attributes, or additional datasets. These changes allowed me to reach the same end goals, just in a way that better fit the data that I had available to me.

Ryan Malone
CSCE567-001

April 21, 2025
Final Project Writeup

## Summary / Visualization Access

People have been buying and selling homes for centuries, and for many Americans, owning real estate remains a key part of achieving the "American Dream." However, like any other market, the real estate market is influenced by larger economic trends. Over the past 20 years, real estate prices have fluctuated due to major events, such as the rapid declines during the 2008 financial crisis and the post-COVID-19 surge in prices when demand was significantly higher than supply.

The primary goal of my visualizations was to look at historical real estate sale information across one of the states, Connecticut, to investigate when prices were high or low, the general effect of different historical market events (such as COVID-19 and the 2008 financial crisis) on the housing market and real estate prices, as well as the number of sales that occurred in different areas. These visualizations would give viewers an idea of what many people look for when they're choosing real estate and how different events can affect real estate transactions. The visualizations I have produced can be broken down into three different categories: an overview, market changes over time, and property type insights. The overview visualization simply shows viewers which towns across the state of Connecticut had the largest number of sales. This information revealed that most individuals purchased property in the southwestern part of Connecticut which may have to do with larger populations of individuals, the level of urbanization, or the proximity to outside urban areas and cities (such as New York City). The market changes over time visualizations explore how sale prices have changed over the past couple of decades and dive into how these sale prices may differ from the assessed value of the same real estate that was sold. This information can give viewers an idea of whether markets tend to overvalue or undervalue properties during certain periods, how prices change over time, and how external events may influence buyer behavior or perceived property value. Lastly, the property type insight visualizations allow users to explore what kinds of real estate were most sold, such as single-family homes or multi-family properties, how the sale prices of different property types differed, and how these trends vary by town or over time.

All of my visualizations are hosted on GitHub pages and can be found at ryanmalonee.github.io/visualization-tools-project. When first accessing the website, you'll be introduced to some background of the project, along with the overview visualization. Users can then navigate through the different pages to see the "market changes over time" visualizations, the "property type insights" visualizations, and some other external resources that may be helpful to understand key events that affected the real estate market. When interacting with the different visualizations, users may hover over the key information to quickly view the tooltip, which contains more information about the data and can be useful in understanding different trends or events that are taking place. Filters are also available on many of the visualizations to find more specific information relating to towns, time periods, and property types.

## Challenges Encountered and Addressed

Throughout my time working on this project, I encountered a few issues when creating my visualizations. The first big issue occurred as I was trying to build the geographic map of Connecticut, showing the number of sales by each town. I wanted the visualization to shade each of the towns in Connecticut, and the darkness of the shading would describe the number of sales (a darker shade meant more sales occurred in that town). When attempting to build this visualization, I realized that Tableau would not render the map the way that I originally intended, instead just showing points across all the towns, rather than shading their boundaries. I was also not able to change the type of map (it was greyed out). To fix this issue, I decided to find an additional dataset that contained all the United States zip codes, as Tableau has built in functionality to work with zip codes as geographic points. After finding a free dataset to use, I imported it into Tableau, related it to the real estate data by the town name (after limiting it to only use Connecticut towns), and was able to then create the map that I wanted.

An additional issue that I ran into occurred in the "Sale Amount Over Time" scatter plot. After setting up the visualization with the sale amount as the y-axis, and the date the sale was recorded on the x-axis, I realized that the visualization was incredibly difficult to read due to some of the outliers that happened (causing most of the sales on the plot to be very condensed in the bottom of the graph). To fix this issue, I created a filter users could use to limit the maximum sale prices. This solution worked very well because it not only allowed the user to filter out the outlier points that were causing the graph to expand, but it also allows users to look at specific sales within a certain range, which can be useful for understanding what price range most sales took place in. After doing this, the second issue took place where I realized that Tableau was automatically aggregating the sale prices, which didn't make sense to me since I had included the serial number (what I thought would be unique to each real estate sale) as a detail. This aggregation made it so that even with maximum sale price set, some points on the scatter plot were still well above it. Upon further investigation, I realized that the serial numbers were *not* unique, and some serial numbers were the same between different towns. To fix this issue, I included the town as a detail as well, making sure that each data point represented an individual serial number and town combination, which was unique to each sale.

Lastly, another issue I ran into was in the "Change In Median Sale Price YoY" visualization where I was not able to create a reference band with the value set to certain years. I realized, after constructing the visualization, that the date value I used was discrete, not continuous. After researching this, I saw that Tableau didn't allow reference bands to be created on discrete variables in the way that I wanted them to. To fix this, I pivoted by using area markers across the years of the x-axis that represented different historical recessions or economic downturns. I formatted the area to match the style of another graph that used reference bands and added labels so that the viewer knew what they were there for.

## Design Decisions

When creating my visualizations, I considered multiple chart types before deciding on the ones that are included in the final version of the project. For example, when considering how to show users which towns had the greatest number of sales, I had considered using a bar chart showing the different towns and their respective number of sales, rather than using a geographic map. After careful consideration, I decided that the map was the best visualization to use to convey this information because most users would find it easier to locate different towns on a map and quickly identify which towns had more/less sales based on their shading. Comparatively, a bar chart would require a large amount of scrolling to find a specific town (since there are many towns in Connecticut), and two bar charts may look very close to the same values, even if one was bigger than the other.

In addition, the "Sale Amounts Over Time" visualization could have been a simple line graph that showed the median, or average, sale prices of each property type across different years, but I thought a scatter plot would make more sense because it has the ability to show individual properties, where *most* properties lie (in pricing), and would not rely on aggregations, or other operations, to display value, rather the actual sale price for each individual real estate transaction was shown.

Lastly, the "Median Sale Price vs. Median Assessed Value" graph had multiple different design options that I thought of, since its main goal was to display the difference between the two values across different years. I thought of just creating a graph that would display the difference alone, across the different years of the x-axis, but after thinking about it, I figured displaying the sale amount, assessed value, *and* the median difference between the two would be the best decision since it would give viewers the most amount of information that they would need to understand and interpret the graph fully. By displaying all the information, viewers could quickly see the sale price and assessed value, but most importantly, how big the difference between the two varied across different years. By quickly looking at the visualization that I ended up with, users can see that one year had a larger difference than another because the bar in between the two lines is longer (or vice versa). In addition, viewers can quickly see which year has the largest median difference in sale amount vs assessed value, or which year had the least median difference.

When considering which color scheme to do, I wanted to choose colors that would highlight the most important information for the user, while also using representative colors. For example, most individuals would view red as a "bad" color, meaning something bad is happening. For this reason, I decided to mark the periods of recession/economic downturn in red, so that users would quickly be able to identify that something bad was happening at a certain period (2008 financial crisis, COVID-19, etc.). In addition, for many of the graphs that had to do with the number of sales, or the sale amounts, I decided to use a green color, since many people associate green with money, prosperity, or something good happening. All the real estate sales involved large amounts of money changing hands, so I figured green would be a good representation of this and would be easily relatable for the viewer. Throughout all the visualizations, CPI-U values were not the most important information displayed, it was simply implemented to give viewers more context about what was happening in the world, so for most of the graphs I used a faint grey line that was still noticeable to the viewer, but didn't immediately grab their attention like the green or red colors. The only exception to this was in the "Sale Amounts Over Time" scatter plot, where the CPI-U line was yellow) since there were so many green points in the background that the grey line would not have been very noticeable to the user. Most of this graph was still green, so by making the line yellow, CPI-U values were still readable, but weren't the primary takeaway from the visualization.

Throughout the entire project, I wanted all my visualizations to display enough information that the viewer could understand what was being displayed, have enough information to gather their own conclusions about the data, but also understand and validate the conclusions that I gathered about the different visualizations. To do this, I relied heavily on the tooltip feature to display all relevant information. In addition to color coding and the axis labels, all my visualizations will display all relevant information when hovered upon. This could be something as simple as displaying a date, so the user knows exactly *when* a real estate transaction occurs, but most of the time more complex information is displayed, such as exact sale amounts, exact assessed values, and the difference between them (in the case of the median sale price vs. median assessed value graph), exact CPI-U % change values from the previous year, the town a transaction occurred in, the property type of the transaction, and the exact number of sales within a category.

Ryan Malone April 21, 2025
CSCE567-001 Final Project Writeup

## Discussion of Future Work

While my project includes a range of visualizations that help explain trends in Connecticut's real estate market, there are several opportunities to expand on this work in the future. One possible expansion could be to incorporate data relating to demographics or income levels into a new visualization. Income and socioeconomic status plays a huge role in what real estate properties someone can afford to buy. By incorporating this data into the project, new visualizations could be produced. For example, a line graph could be constructed to display how the number of sales, or the median value of sales, within a certain area (such as specific towns) differ if they have more individuals with higher income levels. In addition, more specific cost of living data could be introduced into the visualizations to see the difference in the number of sales between different towns with different costs of living (an area with a higher cost of living may have fewer sales). Additionally, I could explore creating a new visualization that maps how long properties typically stay on the market in different areas across the state of Connecticut (some towns/areas may be harder to sell real estate in). This would offer more insight into demand and market conditions across regions, what drives demand, and could not have easily been added to the current dataset that I have obtained. Lastly, the project could be expanded to the entire United States to see how different states or regions differ from another, how sales in different climates differ, etc. There are many possibilities on this topic, and the ideas could be applied to almost any area (assuming there is data available).