Olivia Connelly
Andre Dos Santos
CPSC 4030 Final Project Process Book

Overview and Motivation:

Provide an overview of the project goals and the motivation for it. Consider that this will be read by people who did not see your project proposal.

This project was based on data that is centered around real estate values for properties in Connecticut. We chose this set of data because it took multiple values for each property and classified each property by multiple different factors. Knowing this, our goal for this data was to understand if there were any factors that truly affected the property values of properties in Connecticut. We decided to analyze both the assessed and sale values of each property and to understand if the property type, such as residential, industrial and commercial, affected these values in any significant manner. We also looked to see if the location of these properties would impact these values too. So, to do this, we analyzed the impact of the different towns these properties were located in on the assessed and sale values for each property.

Our goal for this project was to look at this data from the eyes of someone looking to either buy or sell property. If you are looking to buy a property, you want to know what locations and property types will increase the price of a property to buy. Therefore, you would want a tool to help you determine where and what to look for to be more cost effective. If you are looking to sell a property, you want to know how much more you can sell your property for compared to what your assessed value will be. Therefore, you would want a tool that would tell you what property types have a larger gap between assessed and sales values and what property types and towns generally have properties that sell for more.

Our dashboard also allows for the user to understand where each individual property stands compared to other groups of properties and to understand if the selected property may be an outlier and fall outside of the trend or may fall exactly into the expected values.

Overall, our dashboard was designed to be a tool to understand our dataset in a meaningful manner. Without interpretation and comparison shown in the dashboard, there is not a clear meaning of the data, so we sought out to find what are possible impacts on the values of real estate in Colorado.

Related Work:

Anything that inspired you, such as a paper, a website, visualizations we discussed in class, etc.

Both of us have done previous work, internships and classwork, regarding data visualization. We took the experience of what resonated with people and what was discussed in class, specifically interacting with individual visualizations to refine the data and connections shown, to create a website that helped analyze this data as effectively as possible. Seeing in class how you can have one graph that can show every data point that you have and then utilize it to refine other visualizations and show their connections was what we tried to base our dashboard around.

The scatterplot visualizations early on in the semester allowed us to start making a good design layout for our dashboard. Upon seeing it and having it explained in class, we were able to decide that it would best fit the type of data that we had. We had also seen the map types of graphs. However, we were not able to find one that effectively fit and represented our data in a meaningful manner. In class we saw how versatile bar and line charts were and found that they allowed for so many possible interactions. We started with those and found that the bar chart did not properly encompass all of our data. In class, we then saw a version of the connected dot plot. We found that this visualization would allow us to summarize our data more effectively and condensed.

We were also able to get some ideas from the presentations in class. We were able to see how people were using line charts with their interactions. We were struggling to make our line chart interactive and make sense when it changed. We saw that other groups were adding a comparison point to their line charts to allow the user to have a reference point. This allowed us

to shape our line chart interactions for this line chart to make more sense to someone who hasn't seen the backend of the project.

One of our biggest design influences was the discussions of color in class. We found that having overlapping features was useful for our dashboard. However, they all started to blend together. When we saw the different types of colors to be used we were able to decide that leaning towards bright contrasting colors would be effective for us. We wanted to make sure that our data points wouldn't all blend in together so this allowed for a very obvious difference.

Overall, the majority of our ideas for this dashboard came from previous experiences and what we saw demonstrated in class. Class was definitely our most important resource as we were able to get feedback on what was not effective and see real world examples of what types of visualizations were effective.

Questions:

What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

When we started this project, our first questions when making our prototype were what factors are available that can interact with our data? This outlined the goal of our prototype which was to answer this question. After we had created the prototype, we had realized that there were two apparent factors that could impact the assessed and sale values of the properties. So, our questions now shifted towards, how can we show the impact of these factors, which were property type and location. We first tried simple bar charts and line charts which didn't seem effective so we shifted towards answering, how do these factors impact the assessed and sale value of the properties?

This single question had us asking, what graphs would be able to even encompass data like this in a visually appealing and understandable manner, and, how do the data points compare to the trends shown? This sculpted our focus and directed us to start using averages to create these new visualizations. Once we had visualizations that then showed us these comparisons, we were finally able to ask, how does each data point individually compare to averages and how are they impacted by location and property type? We were able to answer this final question by then using the dashboard as intended.

Data:

Source: https://catalog.data.gov/dataset/real-estate-sales-2001-2018

Overview:

The data utilized for this analysis consists of a pre-collected, static dataset of property sales, sourced from a json file in our repository. This data is structured as a collection of real estate transactions, which comes from public Connecticut property sales records from the year 2021. The visualization application loads this data asynchronously using the d3.json method.

Loading:

First, the code is extracted from "Real_Estate_Sales_2001-2023_GL.csv" using a python script that allows us to filter by town, year, property type, etc. Once the filter is selected (if any), the data is then loaded into a json file which our dashboard platform can read. For this project, we set the loader's filter to 2021 and got rid of the extra data such as coordinates and assessor notes. The remaining data was then loaded into the json file which this dashboard pulls from.

Data Cleanup and Outlier Filtering:

Before being used in the visualizations, the raw data undergoes a strict cleanup and filtering process (implemented in the cleanData function) to ensure data integrity and mitigate the impact of extreme outliers. The key cleaning steps are as follows:

1. Type Conversion: The financial and ratio fields (sale_amount, assessed_value, and sales_ratio) are explicitly converted to numeric data types to ensure accurate mathematical operations.

2. Minimum Sale Filter: Records with missing sales data or a sale_amount less than $10,000 are removed, which excludes non-market sales and administrative transfers.
3. Sales Ratio Outlier Removal: To eliminate records where the assessed value and sale price are drastically misaligned (often indicating non-fair-market transfers or data entry errors), records are filtered based on the sales_ratio. Any record with a sales_ratio less than 0.05 (5%) or greater than 5.0 (500%) is excluded from the analysis.

Finally, for performance optimization on large datasets, the application includes a sampling mechanism to limit the active data points to a maximum. In this case, we allowed 100,000 records before passing the data to the charting functions (but there are only 60,000 for our entire data set).

Data Processing for Visualization

Following the initial cleanup, the data is aggregated and processed further by the individual chart scripts (for example, scatterPlot.js, propertyTypeDumbbell.js, townComparisonChart.js) using the D3.js library to extract key metrics.

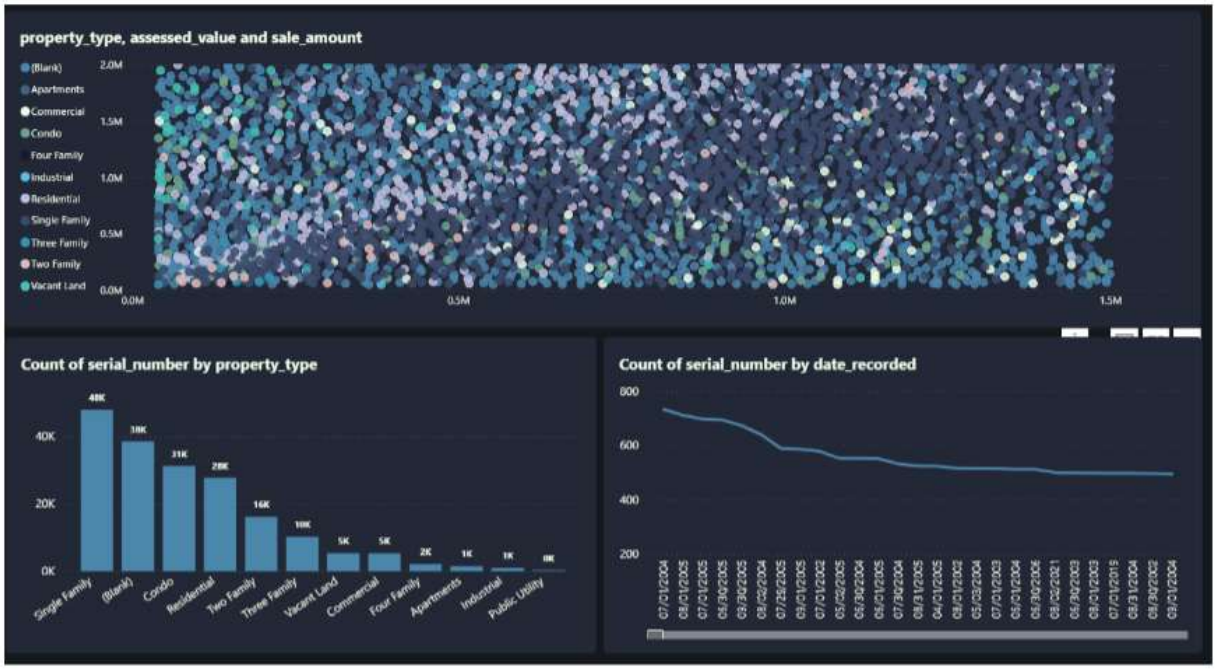
- Property Type Aggregation: For the Dumbbell Chart, the cleaned data is grouped by property_type to calculate the average assessed_value and average sale_amount for each category.
- Town Comparison Aggregation: For the Town Comparison Chart, sales are grouped into bins based on their assessed_value (with a bin size of $200,000) to calculate the average sale_amount within each bin. This prepares the data for trend analysis across assessed value ranges.

- Visualization Filtering and State: The charting components also manage user interaction, such as filtering data based on a selected property_type in the scatter plot, and handling the selection of a single sale (selectedSale) to highlight its values across all visualizations.

Exploratory Data Analysis:

What visualizations did you use to initially look at your data? What insights did you gain? How did these insights inform your design?

Initially, we looked at much simpler visualizations for our dashboard. Below is the original prototype that we were starting with. We were initially aiming at having a general color theme, simple graphs that required scrolling and that were easily implemented.



After creating this prototype, analyzing it and getting feedback, we found that these were not effective visualizations. We found that these visualizations did not convey any meaningful information that would be useful to help one understand the data.

First, we focused on the scatterplot. We decided that the scatterplot needed different colors and had to be refined and condensed to show meaningful trends and data groupings. Once this was complete, we were able to start seeing what factors were actually important to the data.

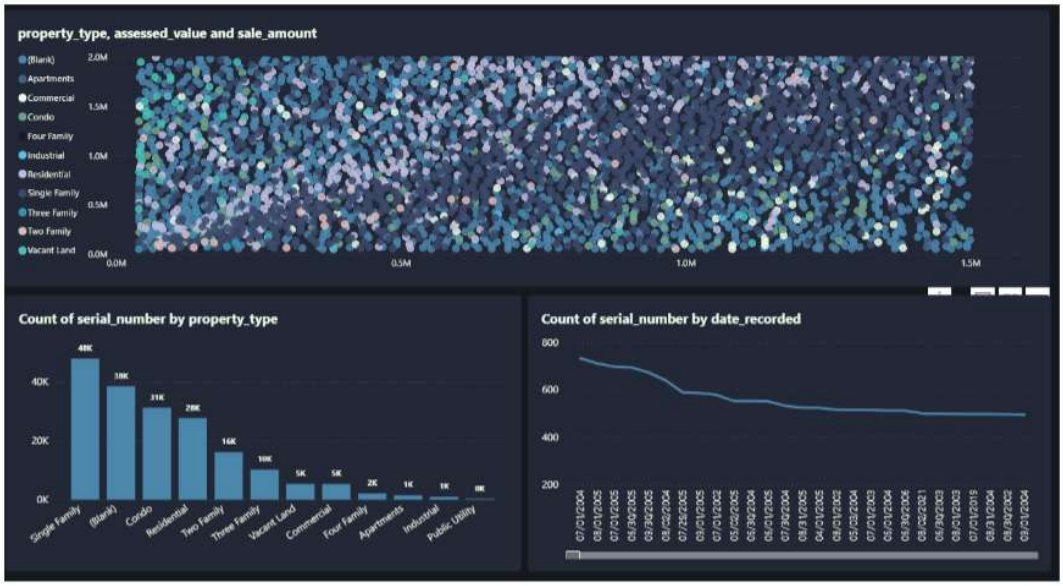
Then, we focused on the bar chart. It didn't seem to tell us anything meaningful by just measuring counts. We then looked to see what other forms of the data would be helpful and found that averages told a lot more than counts. With this, a bar chart no longer worked. We switched to a dot chart which still didn't seem effective. Finally, we moved to a dot and line chart which made the information much easier to understand and lay out without being too crowded or overwhelming.

Finally, we looked at the line chart. Our issues with this chart were very similar to those of the bar chart. However, one glaring issue was that the data just didn't fit into the chart without having the need for scrolling. We looked at a way to condense the line chart which was to go by averages of the assessed and sales values instead of counts. We then wanted to make the line chart more in-depth so we added a way to interact with it so that it would add another line for towns.

Design Evolution:

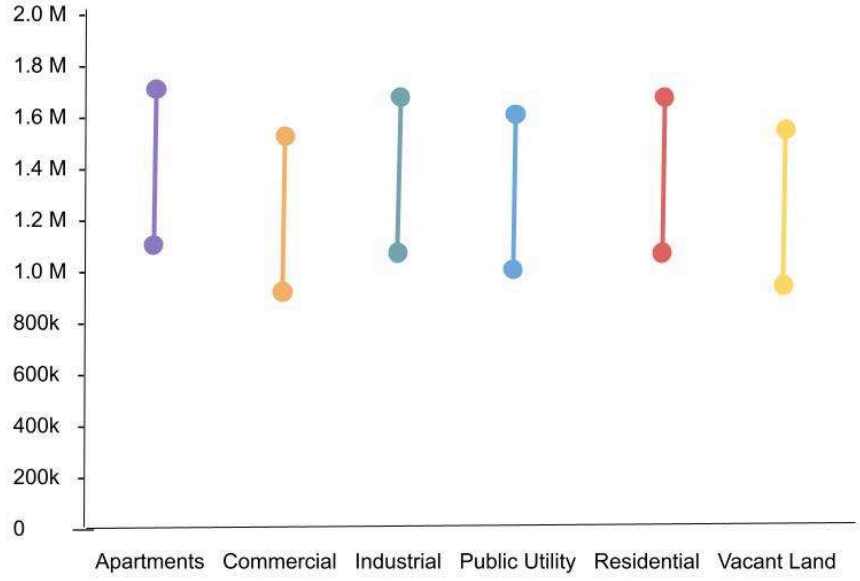
What are the different visualizations you considered? Justify the design decisions you made using the perceptual and design principles you learned in the course. Did you deviate from your proposal?

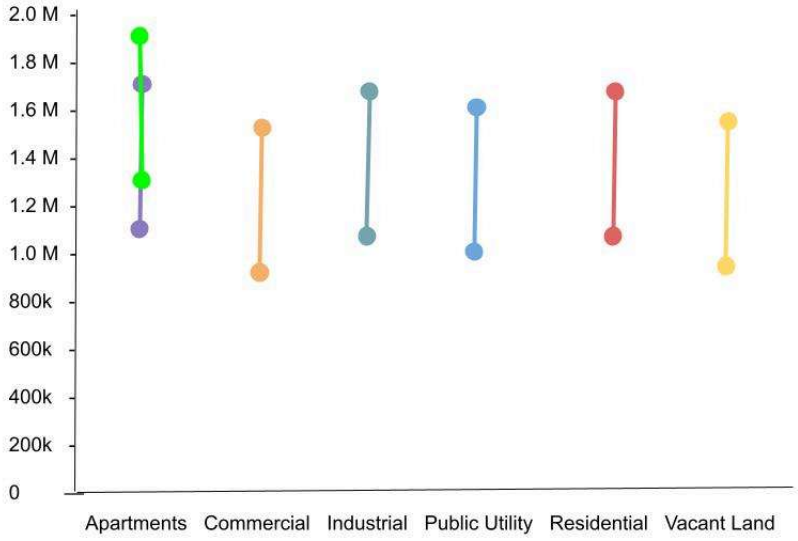
Below are the first prototype visualizations that we considered. We decided to change these visualizations for multiple reasons, most of which came from principle during class discussions. One of the main reasons was the colors in the prototype. The colors all blended together when there was a need for contrast to see trends especially in the scatterplot. The visualizations were also not intuitive so we made them more complex with their implementation but more visually appealing.



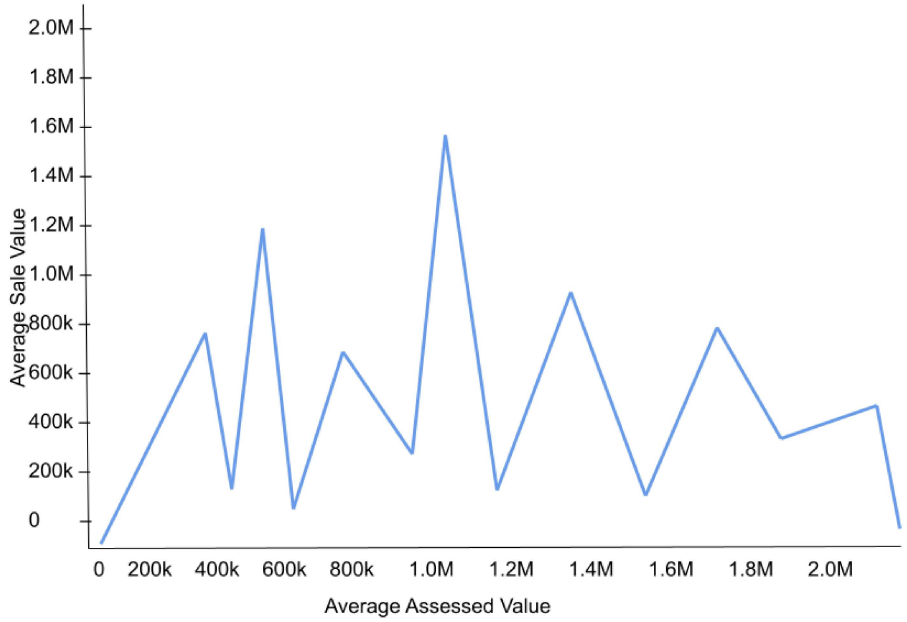
Below is our next iteration of the project. We were able to get better color selections to differentiate property types and much more condensed graphs. We also decided to add a graph to show the raw assessed value. However, this graph did not stay for very long.

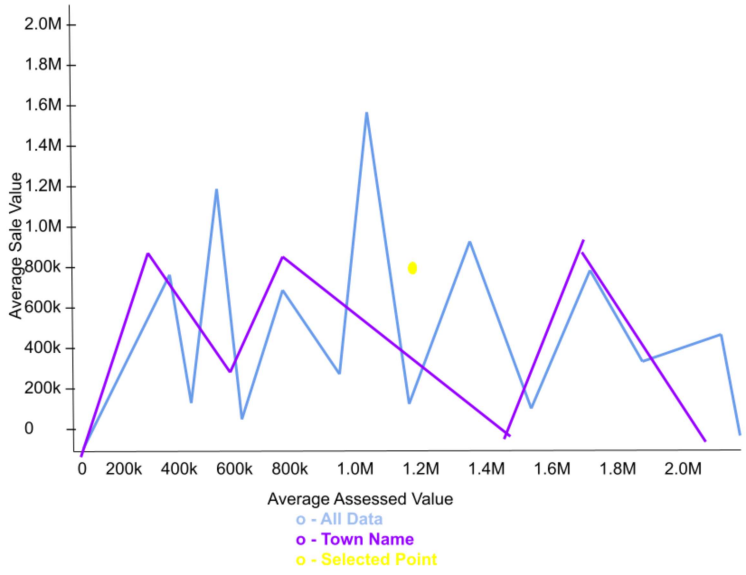
After these first dashboards, we then went on to sketching out graphs. The sketch below is the graph we intended to replace the bottom left and bottom middle visualizations. Our goal was to condense the analysis of the assessed and sales values by property types into one graph instead of the two as seen above. This made it less redundant. The graph(s) below also show how you could then interact to compare the point values to the average values of each property type. This next iteration was much more in depth than the previous designs.





The designs below were intended to replace the bottom right graph. We found that a line chart based on the average values was much easier to read and understand. We also wanted to create a design where this graph would then be interactable. So, we added the ability for it to have a comparison to the town values. This allowed for further data interpretation.



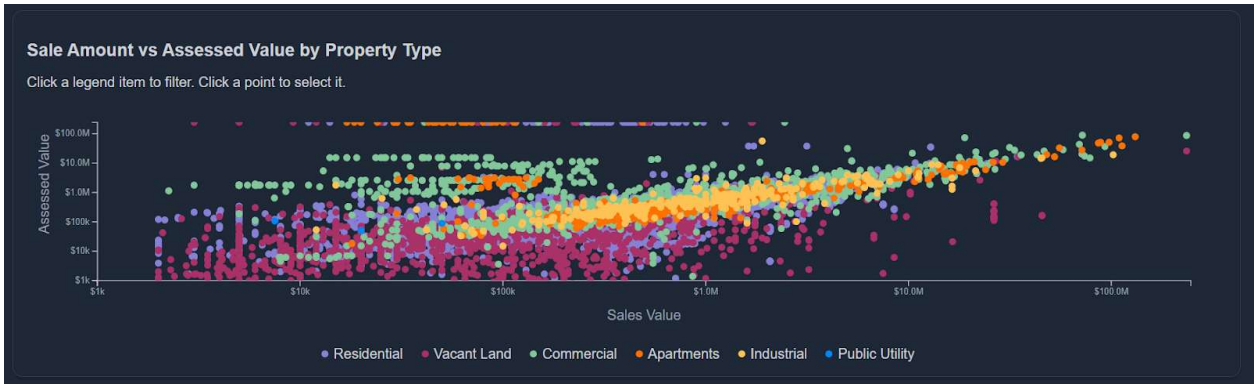


Overall, our final design diverted almost entirely from the original prototype. We kept the idea of the scatterplot as we continued to design the dashboard but worked to polish and refine every aspect. Overall, our later designs were able to answer our questions much more effectively.

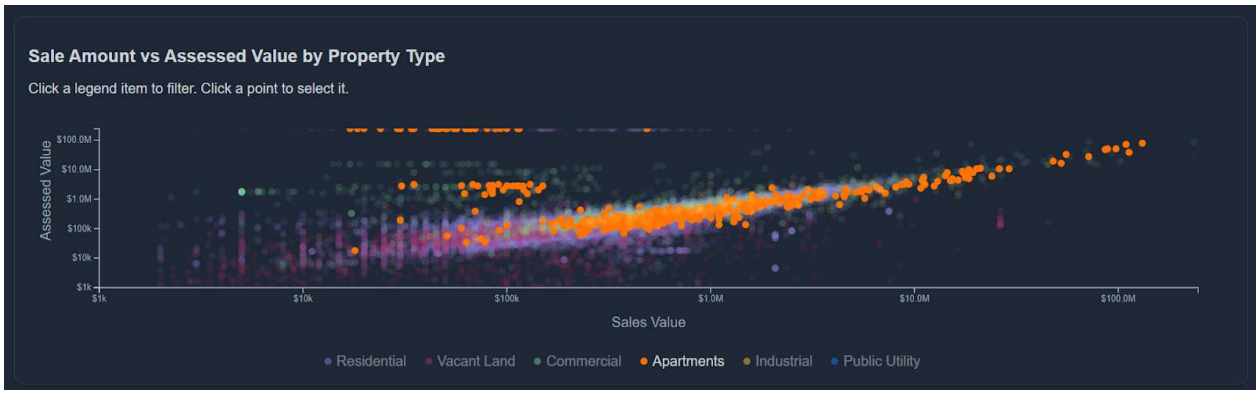
Implementation:

Describe the intent and functionality of the interactive visualizations you implemented. Provide clear and well-referenced images showing the key design and interaction elements.

First, our database starts with a scatterplot of all the available data points of properties. This is meant to give the users a full view of the trends of all of the points together.



We then offer the option for the user to select a property type from the legend below to filter the scatterplot. This allows for the user to see trends that are specific to a property type and compare it to other property types or the data as a whole as shown below. The highlighting and dimming of colors makes it easier on the eyes to see these possible trends and any outliers.



You can de-select any of these property types to return to the original scatterplot and then select an individual point as shown below. When you select a point, it will display all of the information regarding this property and will adjust all of the other graphs in the dashboard. This helps to filter and compare points to the rest of the data.

Below is a graph that shows the average assessed and sales values by the property type. This image below is how the graph will appear when you have not selected a point in the scatterplot. It is a general display of all of the data when it is separated and average.

When a data point is selected, the graph will change to look similar to the graph below. It will then display the assessed and sales value for the selected point on top of the property type that applies to the point. The box displayed on the scatterplot allows the user to reference the exact assessed and sales values that are being displayed for the point without crowding the screen too much.

Below is a line chart that shows the average assessed and average sales values for all of the data. When a point is not selected, the graph will display as below as a comparison point is not yet chosen.

Below is how the line chart will appear when a point is selected. When a point is selected, an additional line will appear on the chart showing the average assessed and sales values for the applicable town for the point. The point itself will also appear on the graph unless it is an extreme outlier to the averages.

Evaluation:

What did you learn about the data by using your visualizations? How did you answer your questions? How well does your visualization work, and how could you further improve it?

From using our visualizations, we were able to learn more about how important property type and location is in our specific data set. We were also able to answer our questions gradually with each step we took. Overall, by trial and error with each new visualization, we were able to gather a greater understanding of the data. This allowed us to determine what types of trends the data was following and the best ways to dig deeper into the data to make use of it.

Overall, our visualizations are very reliable and work to show interactions very well. Every button is functioning and the displays are concise. However, there may be room to adjust the sizes of our graphs. Unfortunately, this was a lasting issue with how big the values of the data were. Overall, having the dashboard run faster and have better sizing would be possible ways to further improve it.