CS 5630: Final Project

Vista Marston & Chris Marston
Fall 2022

Process Book

Table of Contents

Initial Design Process

Process Book

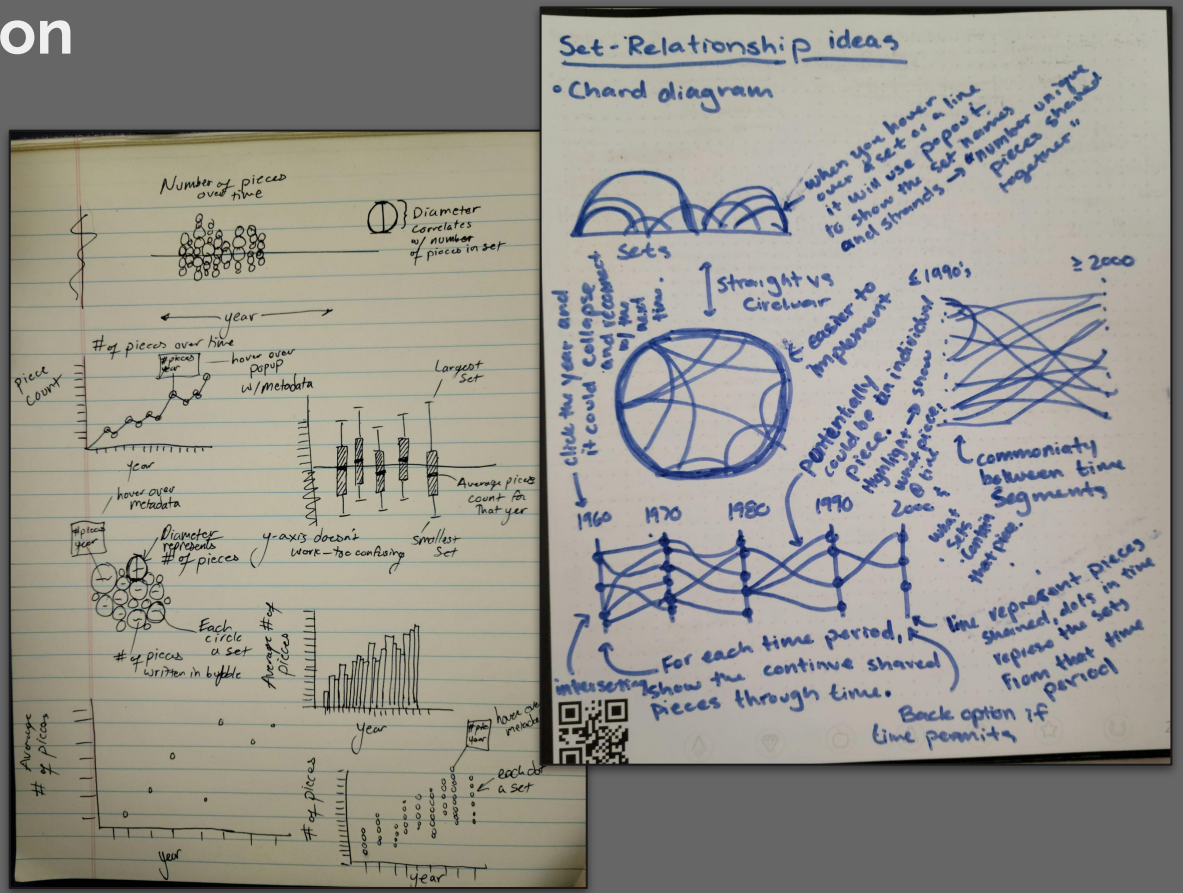
Initial Project Proposal

Summary

Our initial intent with this project was to generate visualizations that allow a user to explore how Legos have changed over time as well as how sets and pieces are related. Rebrickable, a site for posting custom Lego builds, maintains a database of all extant Lego sets and pieces. From this data we planned to generate two visualizations: a chord diagram that shows the relationships between sets, themes, or pieces and a scatter plot that shows how the number of pieces has changed over time.

Design Ideation

Some initial design sketches.



Initial Design Proposal

https://LegoVisualizerOverTime.com

2-Day Elephant & J... | ShareBox Log In | AI courses | Scholarships | Python | ML excerises | Internships | Everyt... | First Time Visiting L... | School | Other bookmarks

Lego Visualizer

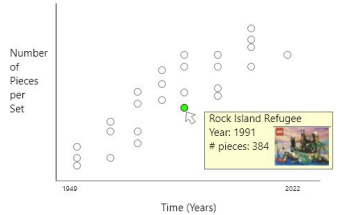
Introduction

Here is One of us (Chris) spent a considerable amount of time building with Legos when he was younger. Although he hasn't built anything with Legos in some time, he remains intrigued by them and is interested in learning more about them. In terms of the whole group, however, we both feel that the data set we found that details the different parts available and which pieces are in each set with their quantities, among other information, will be a lot of fun to work with. It provides a lot of interesting information on Lego sets and we initially had some difficulty narrowing our focus for this project because of the large number of interesting things we could learn about Legos from this data set.

Overall, our primary motivation for this particular project is that we feel that the data set we have to work with gives us a lot of opportunity to learn. In addition to this, we feel that it would be fun to build a site that allows you to explore Lego sets and/or pieces and trends regarding Lego's publication of different sets and pieces.

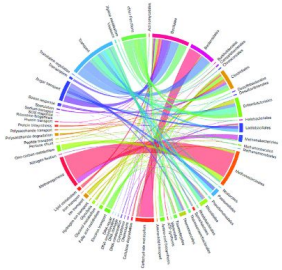
Pieces Over Time

Our intent with this visualization is to show any trends in the number of pieces Lego includes in their sets as well as information about each set. The visualization as described thus far will encode any trends that might exist in the data and we will encode the set metadata using interactivity. The user will be able to hover over each of the dots on the visualization whereupon a textbox with information about the set represented by that dot will be displayed. The metadata we plan to include will be the number of pieces in that set, the name of the set, and the year that set was published.



Set Relationships

We chose a Chord diagram for this purpose because this part of the project does not contribute directly to the overall question that we are trying to answer.



Updated Project Proposal

Change of Design

Once we began processing the data from the Rebrickable database, it immediately became clear that we were dealing with far more data than we had anticipated. Our initial design was conceived under the assumption that we would have only a few hundred sets and a few dozen themes to visualize. Since 1949, the first year in the Rebrickable data, Lego has published thousands of sets under a few hundred themes. Far more than our design could effectively visualize.

After reevaluating the data, we decided that a redesign was necessary.

Updated Project Proposal

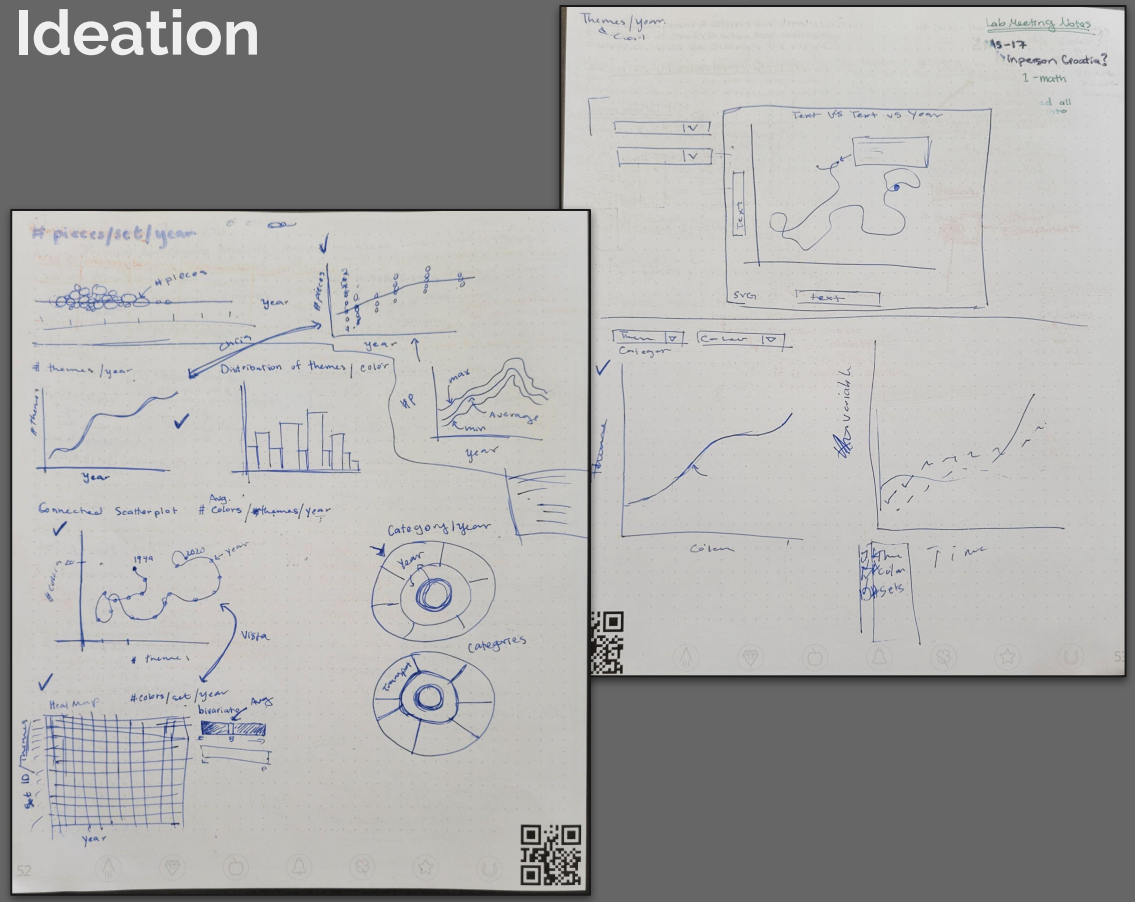
Due to the size of the dataset, we needed visualizations that would be able to effectively display a large amount of data in a concise and understandable format. To accomplish this, we chose to construct a scrolling story + exploration based website where the user can get a high-level overview of changes in Legos over time, but also explore deeper with interactivity. Additionally, we opted to focus solely on how variables have changed over time instead of how sets relate to each other through pieces.

The new design consists of four separate visualizations, each using a basic type of visualization to ensure that the large amounts of data are able to be effectively visualized.

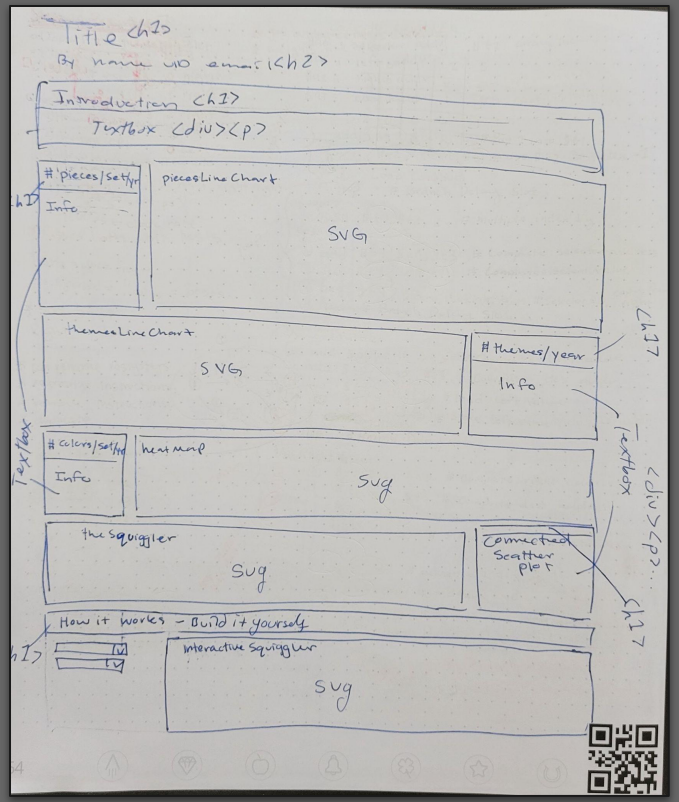
- Our original scatter plot depicting each set, its piece count, and the year it was published.
- A line plot with time on the x-axis and the number of themes, unique colors, average pieces per set, and number of sets published – all per year – on the y-axis.
- An connected scatter plot that allows the user to scroll through a selection of relationships between avg. number of pieces, number of themes, and avg. number of unique colors – all per year.
- A heat map that shows how avg. unique colors vs piece count and the frequency of each color – both per year – have changed over time.

Redesign Ideation

Some redesign sketches.



Updated Proposal Diagram



Process Book

Visualization Build Process

Data Processing

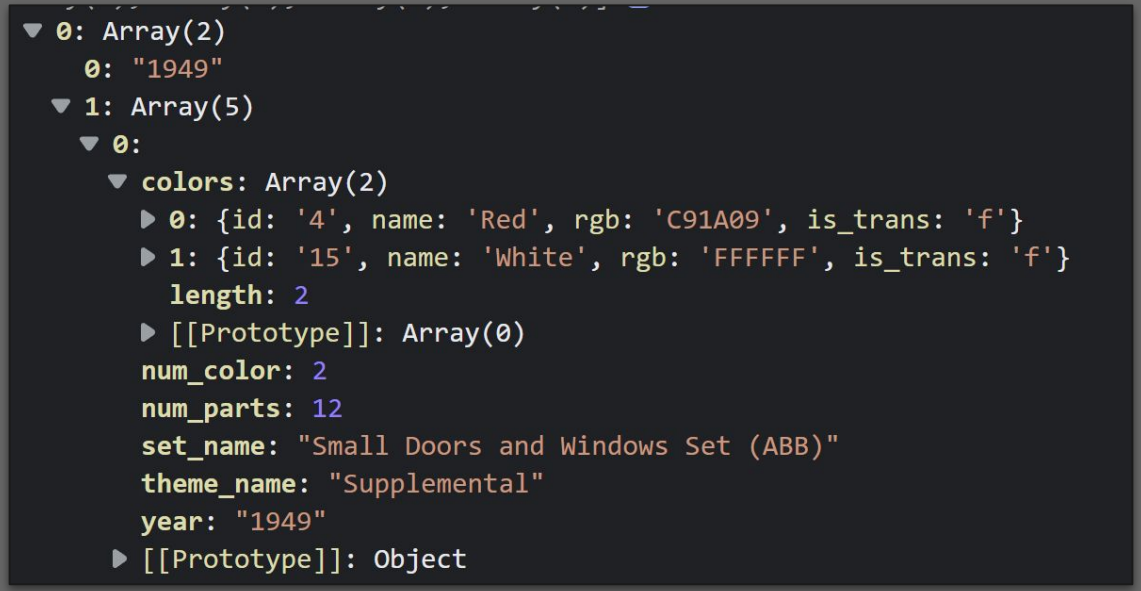
Initially, because the data was coming directly from a database in the form of CSV files – one for each database table – we expected to do little to no data processing. However, it quickly became clear that we would need to join the data from the tables together.

This issue did not pose to large a challenge as we were able to write a script to group the data into an array of objects that could be easily managed with JavaScript. This script runs as part of our website, which leads to slower load times but keeps the original CSV files in the source code for the website.

Additionally, when designing our data processing script, we chose to only process the data from the CSV files that would be used in the visualizations on the website. The variables we collated are as follows.

- Each year as an array of objects, each representing a set.
- Each set object contains the following variables,
  - Year published
  - Set name
  - Theme name
  - Piece count
  - Number of unique colors
  - An array of color objects,
    - Color ID
    - Color name
    - Miscellaneous unused variables

Data Object Structure

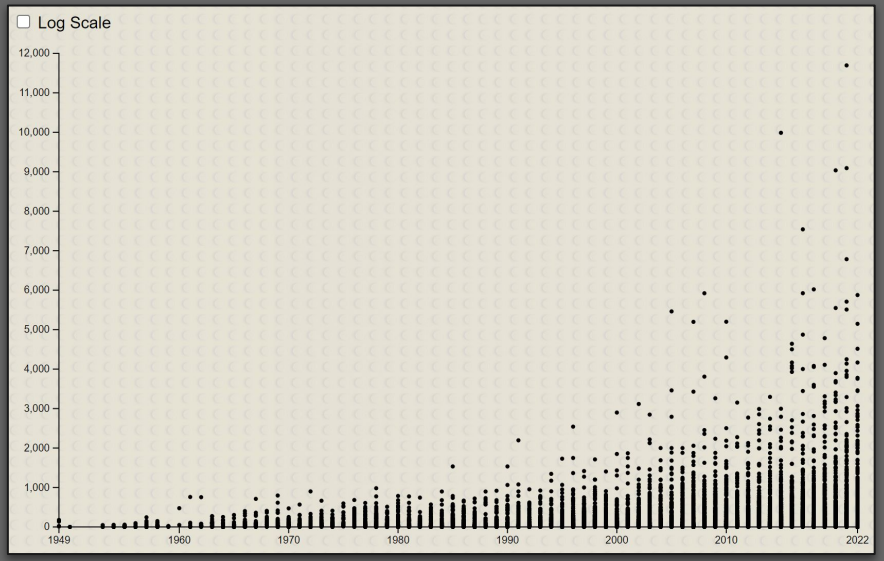


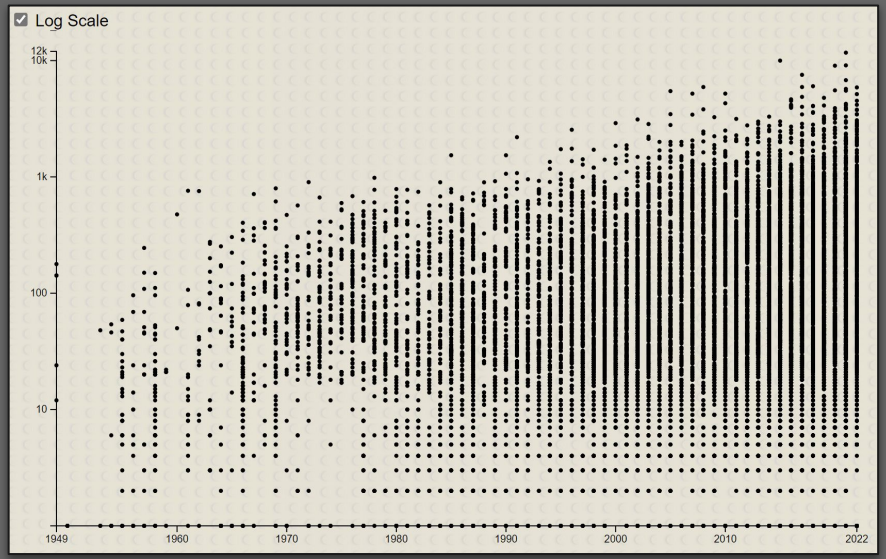
Piece Count Scatter Plot

Upon implementation of the scatter plot, we discovered that the linear scale we used for the y-axis resulted the compression of a large amount of the data due to a few outliers – sets with a very large piece count. To remedy this, we added a toggle to allow the user to switch the plot's scale between linear and logarithmic.

Additionally, we found that simply viewing the chart was not enough. There were several outliers that one might want to explore, so we added tooltips to the dots on the visualization to allow the user to view the set name, piece count, and year published in a popup tooltip box.

Linear & Log Scales of the Pieces Scatter Plot





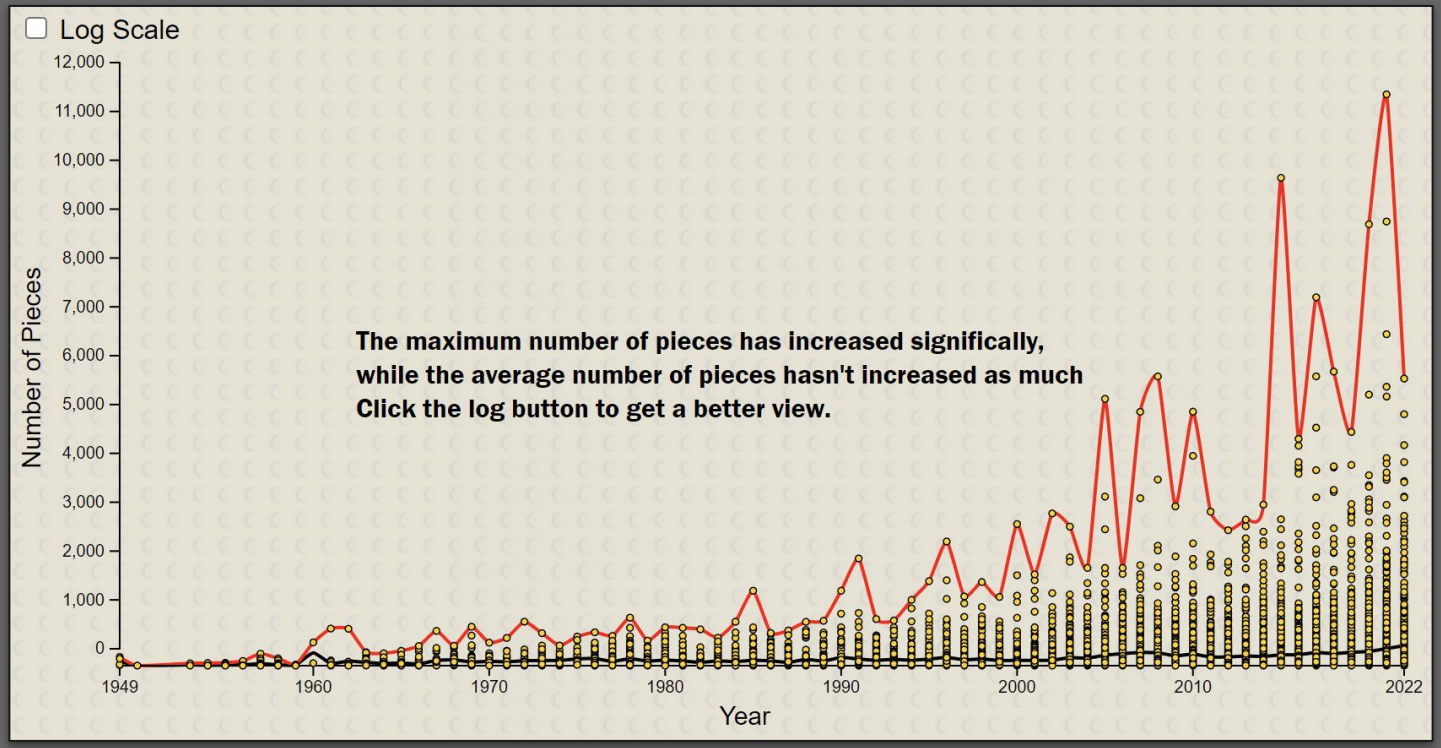
Piece Count Scatter Plot, Second Update

Despite the improvements afforded by the addition of a log scale, the Piece Count Scatter Plot was still a bit difficult to interpret. The primary issue was that there were additional trends visible in the dataset that couldn't be seen in the simple scatter plot alone. Specifically, we wanted to be able to show that while the average number of pieces per set had changed only modestly over time, the maximum number of pieces in a set had increased significantly – Lego has kept their set sizes about the same while offering a few increasingly large sets each year.

To help better display this trend, we added a couple of lines that draw onto the scatter plot after it has rendered. One line, the red one, shows the maximum number of pieces offered in a set each year – essentially, the largest set – while the black line shows the average piece count of sets for each year. These lines plot on both the linear and log scale versions of the visualization and make it much easier to see the trend that we want to show – they make it explicit.

In addition to this, we added a simple storytelling element to the visualization in the form of overlaid text, that explains these trends to the user and prompts them to switch between the different views (linear vs log) in the visualization.

Updated Piece Count Scatter Plot, Second Update



Piece Count Scatter Plot, An Idea - Jigger it?

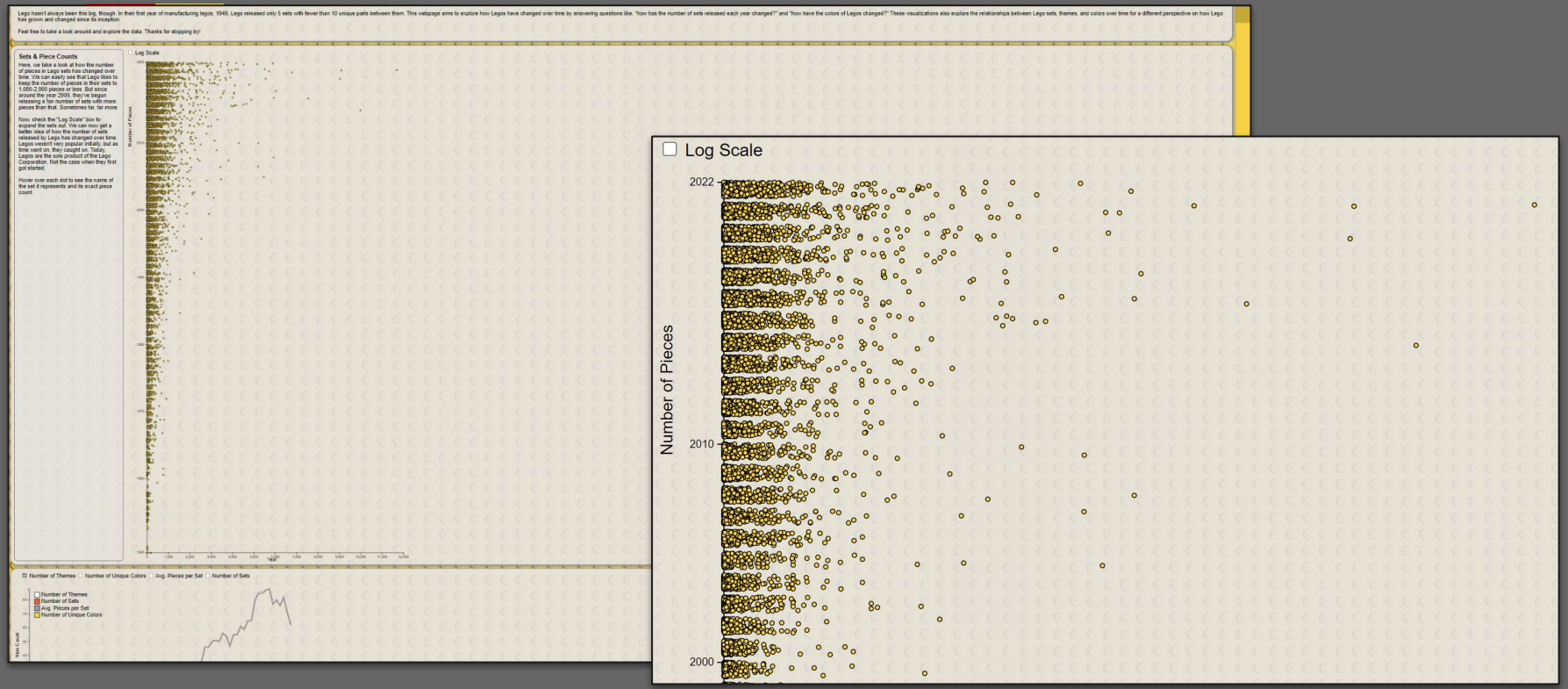
After we built the Piece Count Scatter Plot, it became apparent that not all of the dots (sets) on the plot would be visible or accessible via the hover-over tooltips. The reason for this was that we had not anticipated that there would be as many sets as there were – 17,000 – in the data set.

Because our intent was to show trends in the data and not to make all of the sets available for user interaction, we chose to stick with the scatter plot format for this visualization. However, we did still want to make an attempt to make more of the individual sets visible and interactive (tooltip).

To accomplish this, we tried to add some space between each of the sets using a Jigger Plot format. Unfortunately, this required that each column of dots take up 3-4 times the width that they originally had. This meant that the plot itself had to get 3-4 times wider. Needless to say, it no longer fit on the screen. To compensate, we rotated the plot 90-degrees so that “year” was on the y-axis and “piece count” was on the x-axis, but while this allowed the user to scroll down to see the entire visualization, they couldn’t see it all at once and it took up a very large portion of our webpage. This both obscured the trends we were trying to show and made the plot the primary focus of the webpage.

Because of these complications, we opted to stick with the standard Scatter Plot format instead of the Jigger format so that the entire plot was visible to the user at once and so that it didn’t become the primary focus of the webpage.

Attempted Jigger Plot

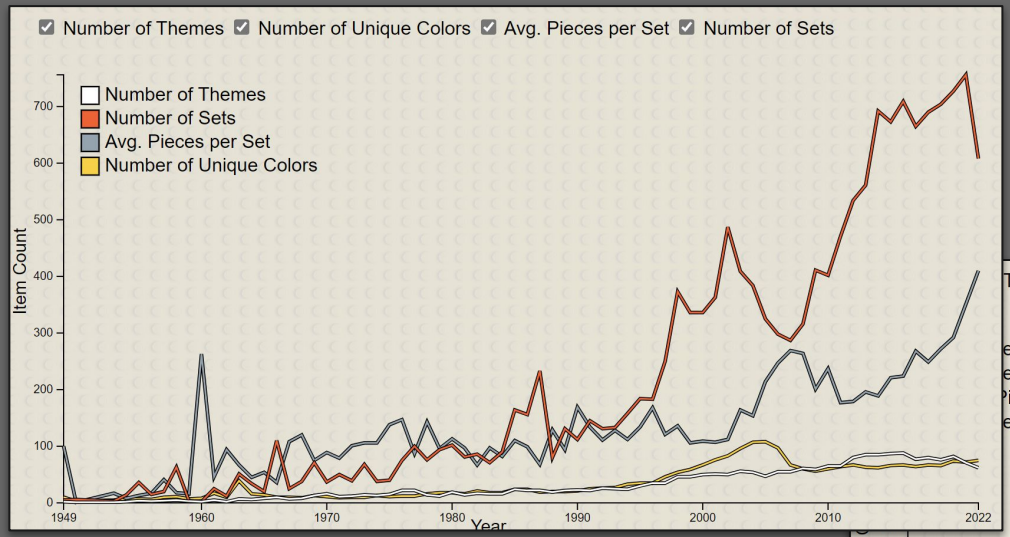


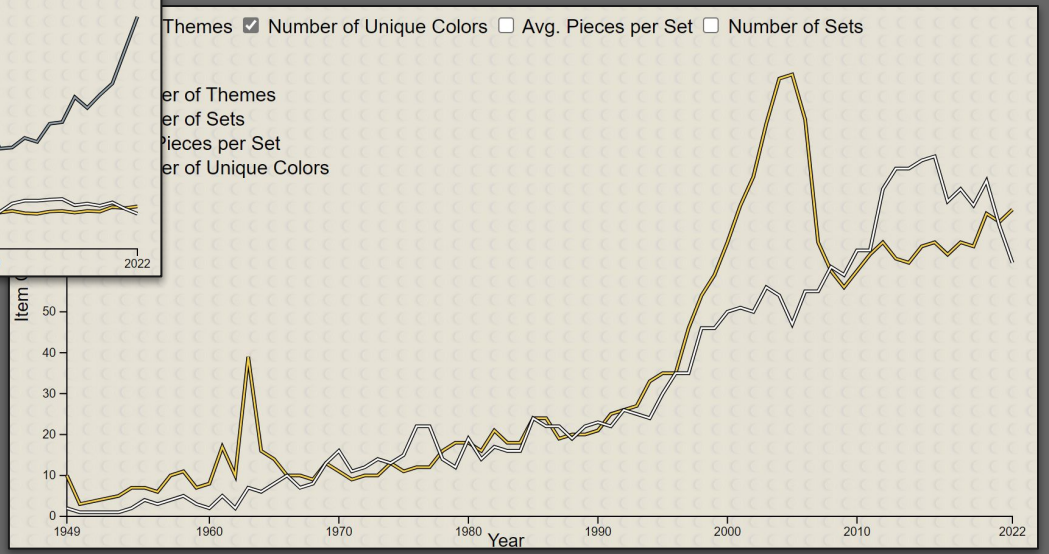
Trends Over Time Visualization

Our original line chart that displayed the changes in themes, colors, average pieces, and sets-published per year was intended to show how different variables changed relative to each other over time. Unfortunately, we failed to consider the fact that the number of pieces per set would more than likely be much larger than the number of sets or themes published each year. This trend also holds true for the number of unique colors used in sets. Because of this, the lines for the number of sets and number of themes become squished when plotted on the same y-scale as the number of pieces and colors.

To remedy this, we chose to add an interactive feature that allows the user to toggle which lines are displayed on the plot. When different selections are made, the y-axis scale adjusts so as to best display the current selection of lines together. This adjustment allows the user to explore the trends in the sets and themes in more detail by adding or removing different lines from the chart and also makes it possible to plot all of the individual lines on the same visualization.

Themes & Sets Only vs All Variables





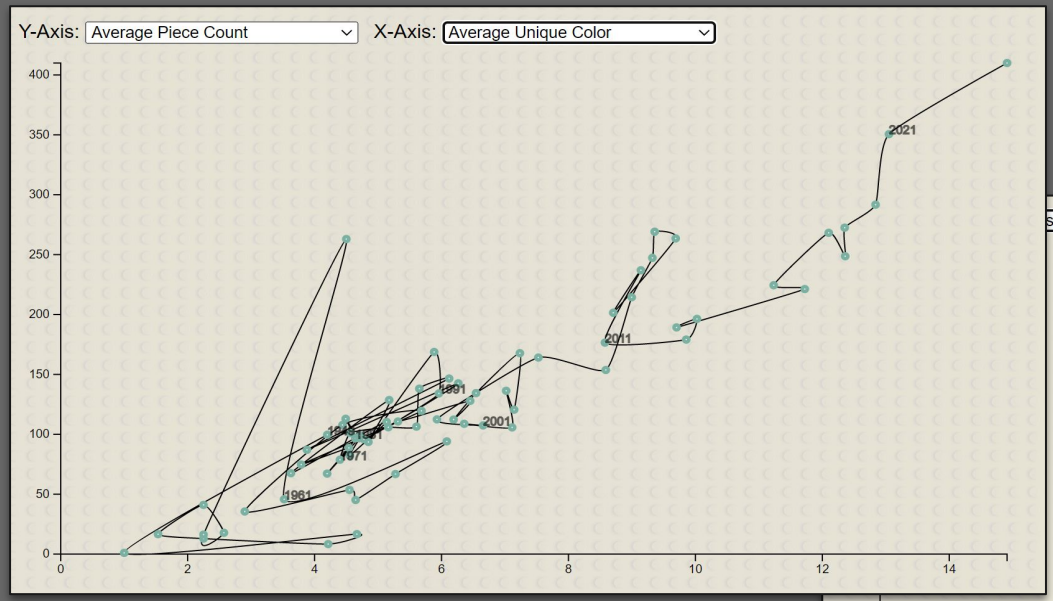
The Squiggler

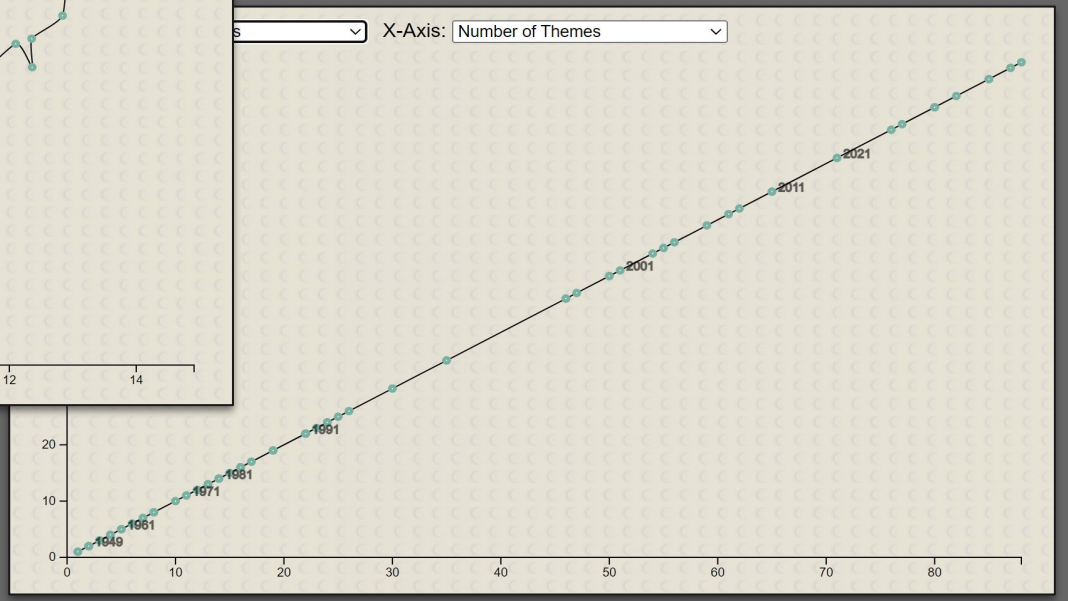
The connected line plot, affectionately nicknamed “The Squiggler,” worked out largely as we had anticipated. Our goal with this visualization was to allow the user to compare different variables with each other so that they could see how the variables changed relative to each other over time. To do this, we initially chose to add two dropdowns that allow the user to select the variables they want to display on the x and y-axes.

While the plot itself came out looking fine, there were a few unintended consequences associated with allowing the user such a high-level of freedom in choosing what is displayed on each axis. The first of these is that if the user selects the same variable for both axes, an uninteresting 45-degree line is plotted. The second is that a few of the combinations of variables produced large “hairballs” in the plot that were both impossible to interpret and unsightly.

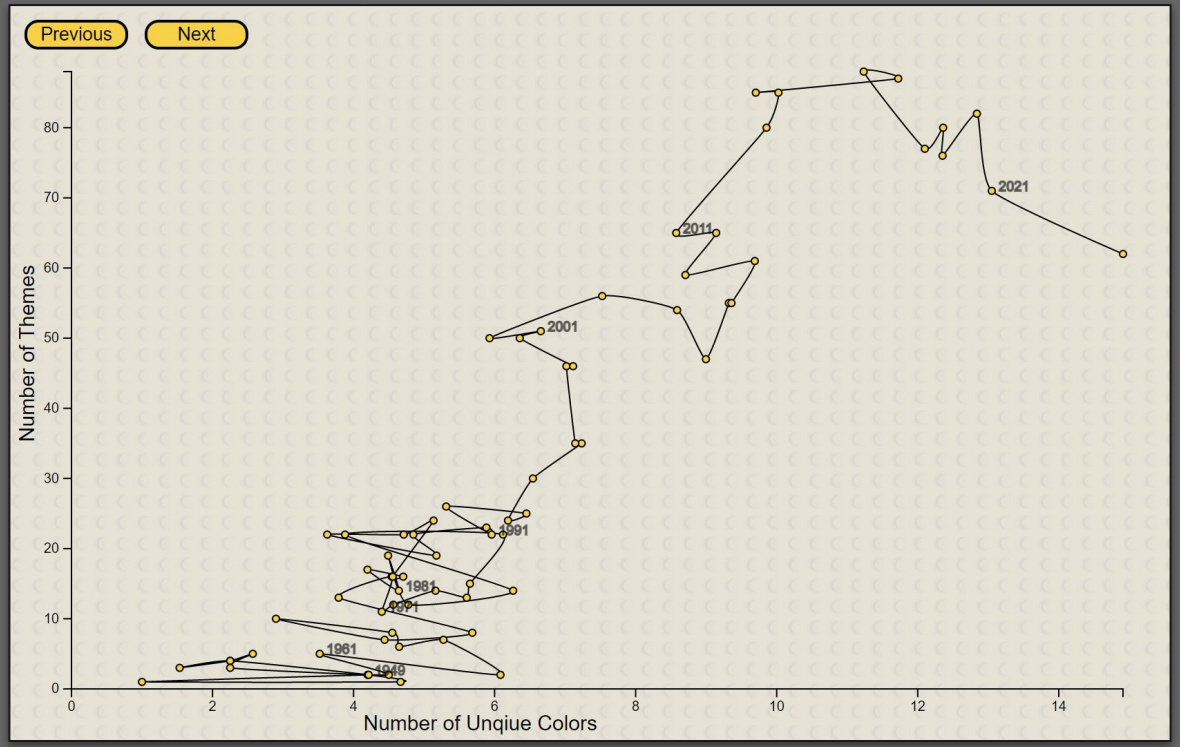
Because of these issues, we chose to replace the dropdown selectors with “previous” and “next” buttons that allow the user to scroll through preselected axis choices. This change eliminated the 45-degree lines and the “hairballs” from the visualization and made the remaining visualizations more informative.

A “Hairball” and a Linear Relationship





Updated Squiggler



The Squiggler, Second Update

After completing the updates to The Squiggler, it became clear that the changes weren't adequate. The "next" and "previous" buttons did not effectively guide the reader through the trends shown in the visualization. Specifically, they weren't descriptive, which left it up to the user to interpret what each of the different relationships were showing.

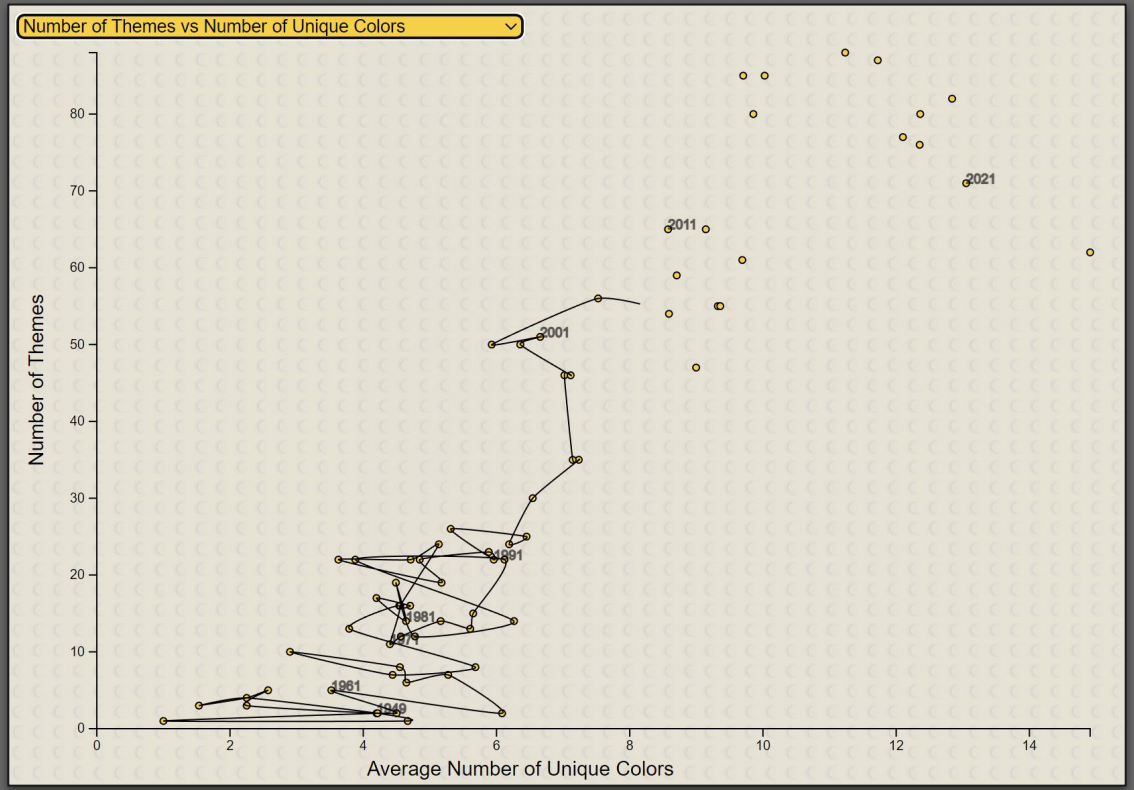
To remedy this, we replaced the "next" and "previous" buttons with a drop down menu from which the user could select from a list of different relationships. This arrangement allowed us to include more descriptive information about the different relationships shown in this visualization and made it easier for the user to navigate between the different relationships.

In addition to the nondescript navigation, we also found that the connected line plots were a bit hard to interpret. Specifically, because parts of them were compact, it made it difficult to see where they started (1949) and where they ended (2022).

To fix this issue, we added a transition that draws the lines between the dots on the visualization in sequential order, from oldest date to newest date when the visualization first loads. This addition helps show the user where the beginning and end of the connected line plots are.

Updated Squiggler, Second Update

Connected line chart being drawn.



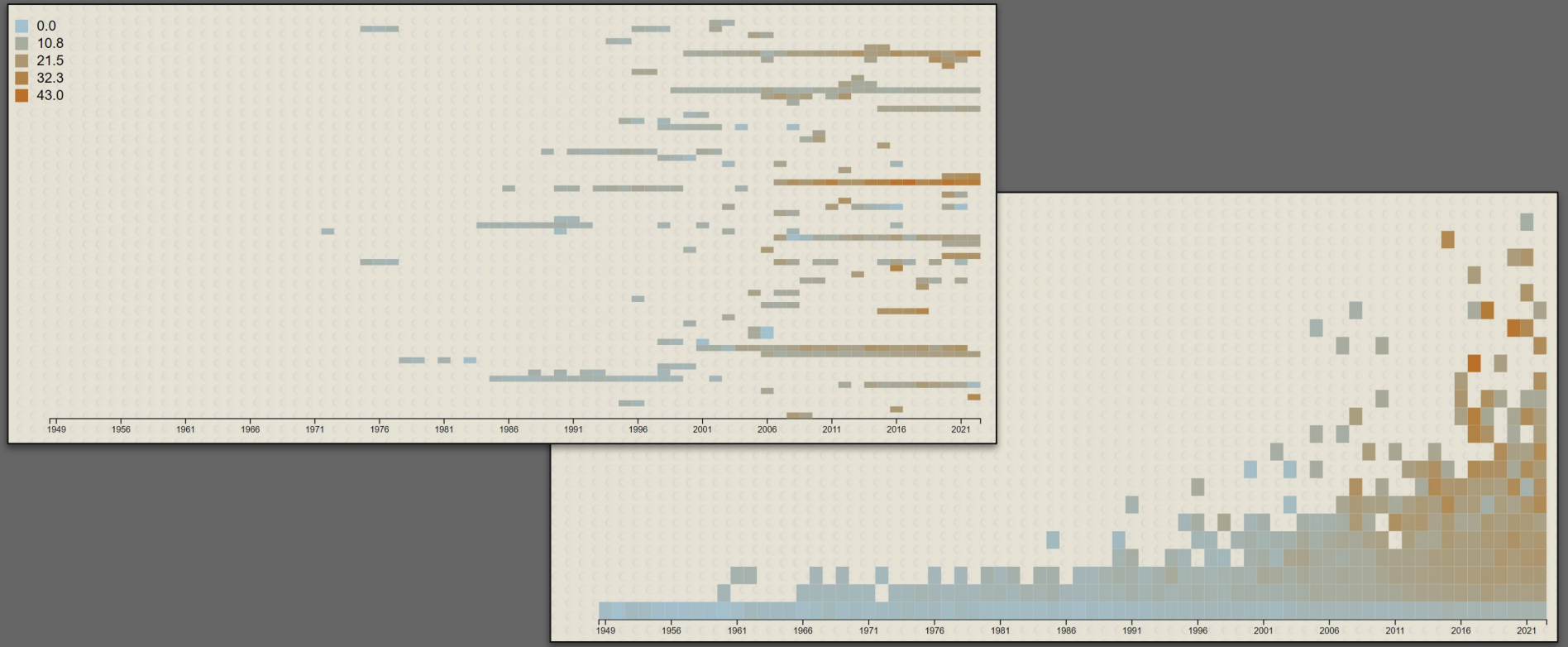
Heatmap

The heatmap visualization also gave us a bit of trouble. We had originally intended to use it to show how the relationships between variables have changed over time, while still being able to show all of the individual sets, pieces, or themes. Since there are so many of them, the heatmap seemed to be a good way to show this. However, when first constructed the visualization, the relationships between the different variables – pieces vs themes for example – weren't strong enough to produce a very strong trend in the visualization. Additionally, because some years don't have elements that fall into certain categories, there were large gaps in the heatmap that made it uninteresting and hard to discern any meaningful trends.

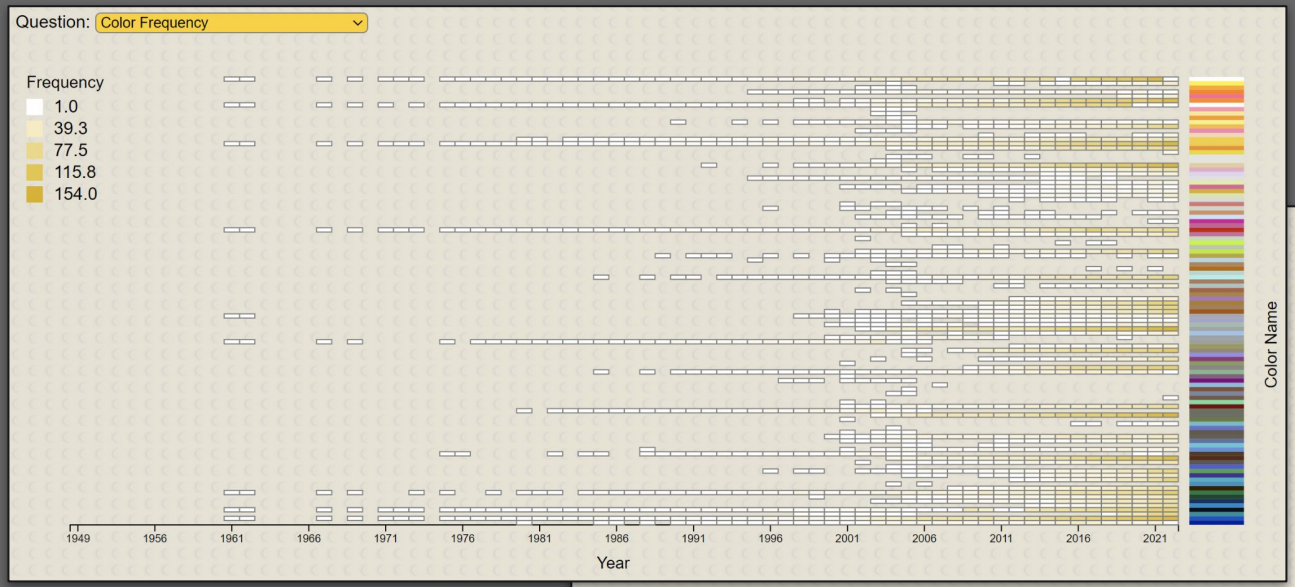
To adjust the heatmap, we selected two different relationships that produced stronger relationships and more compelling visualizations. These two relationships were 1) Frequency of each color used in sets published by Lego over time and 2) the number of unique colors over time relative to the piece count of the sets. We then added a selection dropdown to allow the user to switch between the two visualizations and added tooltips to display metadata for each of the heatmap cells.

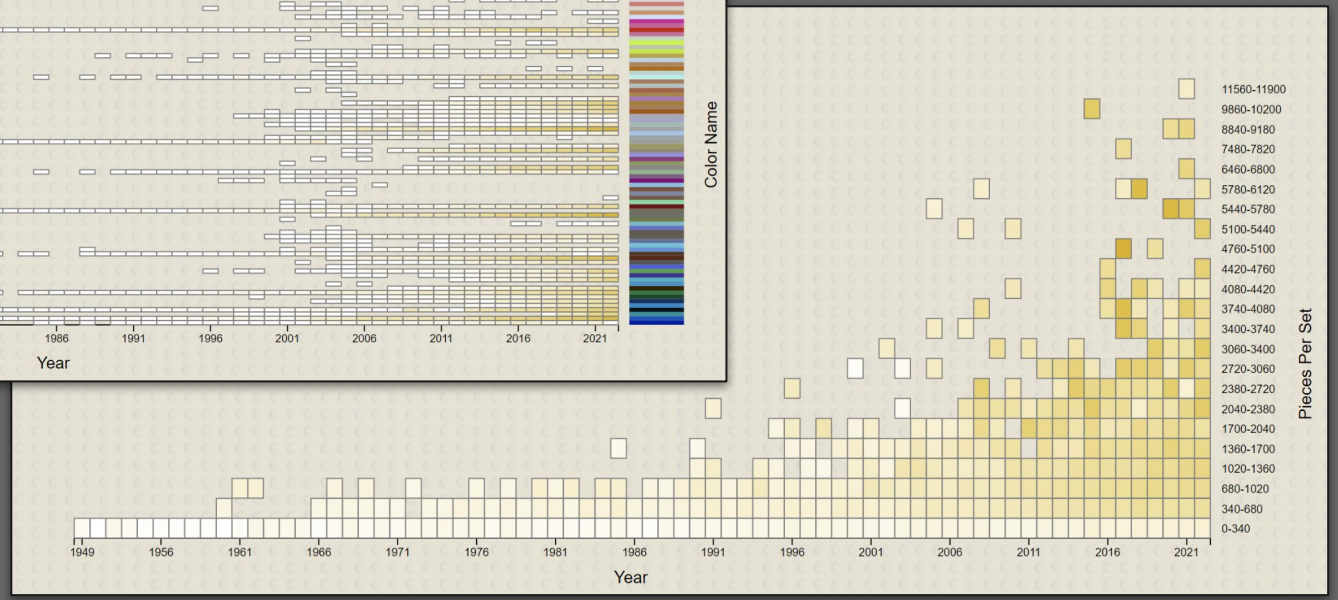
Two additional small adjustments that had to be made were the addition of color swatches next to the Color Frequency Heatmap (1) and the binning of sets in relationship (2) in order to be able to display the data in a concise manner. That is, sets were binned by the their piece count.

Sparse & Compelling Heatmaps



Updated Heatmaps





Process Book

Analysis & Conclusions

Analysis - Visualizations

Our initial evaluation of the Bricklink Lego dataset didn't provide us with much insight into what trends would exist in the data. As we began working through the dataset to determine what would be interesting to explore and what visualizations to put together, we began to see more interesting patterns in the data.

The first of these observations was that the number of sets that Lego has published since they began producing Legos was far higher than we expected it would be. Lego has published around 17,000 sets. This information isn't something we included directly in the visualizations – although we did include it in the text that accompanies them – but it did, as was discussed above, influence the way we designed our visualizations. The Pieces Scatter Plot, which we built before we checked the number of sets (oops!), ended up having dense columns of dots and loads a bit slowly. The modifications discussed above helped compensate for the minor visualization issues associated with this, but we weren't able to solve the lag problem.

Continued on next page…

Analysis - Visualizations

Ultimately, the Pieces Scatter Plot ended up showing some interesting trends. The first of these is simply the sheer number of pieces. While we didn't explicitly display the piece count, when the log scale option is selected, the sheer number of dots on the plot make it clear that Lego has published a very large number of sets.

The second trend was that of the maximum piece count of a single set vs the average number of pieces in all of the sets published in a given year. As is illustrated by the line charts that we plotted over top of the scatter plot, the largest set by piece count published each year has increased rapidly, while the average number of pieces in all sets published each year has remained roughly the same for decades.

Finally, the trend in the number of sets published per year is partially visible in this visualization. It isn't immediately apparent in the linear scale version of the visualization, but when it is expanded to the log scale version, the density of dots on the plot clearly show that Lego has been releasing more and more sets each year. Although the exact behavior of this trend isn't fully clear beyond 1980/90, so this isn't a trend we focused on with this visualization. Instead, this trend is shown explicitly in the visualization just below The Pieces Scatter Plot on the webpage, the Trends Over Time visualization.

Continued on next page…

Analysis - Visualizations

The next trend in the dataset that we found was the relationship between the following four variables:

- (1) Number of Themes
- (2) Number of Sets
- (3) Avg. Pieces per Set
- (4) Number of Unique Colors

The point of interest that stood out in this dataset was that (2) and (3) grew fairly large since 1949, especially (2), while (1) and (4) only grew modestly over the same period. In addition to this, it was interesting to see that these two “groups” were present in the dataset.

Now, that said, while the visualization turned out nicely and does a good job of showing these trends, neither of these trends are particularly insightful. Logically, it makes sense that as a company grows, they will produce more products. Additionally, the trend that (2) and (3) grew quickly while (1) and (4) grew slowly is also no surprise. The number of colors and themes will naturally grow slowly relative to the number of sets released and the number of pieces in those sets.

Nonetheless, this visualization is both effective and interesting.

Continued on next page…

Analysis - Visualizations

Our Connected Scatterplot (aka. The Squiggler) proved to be our toughest visualization. Our intent with this visualization was to to show the relationship between the variables in the Trends Over Time plot (discussed above) in a unique and interesting way.

The visualization displays following variables on the x and y-axes in various arrangements as selected by the user:

- Average Number of Pieces
- Number of Themes
- Average Number of Unique Colors

The first sticking point with this visualization is that we were only able to include three of the four variables that are plotted on the Trends Over Time plot. This was discussed above, but in short, the Number of Sets variable had to be dropped entirely along with some arrangements of the remaining three variables because they generated “hairballs.” Although this change produced a bit of an inconsistency between The Squiggler and the Trends Over Time visualization, it did improve the ability of the visualization to clearly convey information.

Continued on next page…

Analysis - Visualizations

Small inconsistencies aside, this visualization still doesn't show anything new or interesting that isn't already displayed in the Trends Over Time visualization. Aside from the baby "hairballs" that are still present between 1949 and ~1990, the trends between the variables are clear, but those trends aren't different trends from that of the Trends Over Time visualization. And unfortunately, this format is less effective at showing those trends than the line chart used in the Trends Over Time visualization.

So while this plot is unique in and of itself, it isn't the most effective way to present this data. However, we did choose to leave it in the final version of the webpage because it does provide a slightly different perspective on the trends shown in the Trends Over Time visualization and also adds a point of interest to the webpage.

Continued on next page…

Analysis - Visualizations

Finally, we have the Heatmap at the bottom of our webpage. This visualization shows two interesting trends. The first is in the Average Number of Unique Colors vs the Set Piece Count. As the number of pieces in Lego sets increases, the number of unique colors also increases. This seems pretty intuitive, and it is, but Lego could have easily introduced more unique colors over the years without making sets larger. So this helps us see what actually happened.

The second trend that is made visible by the Heatmap is the way color usage has changed over time. This isn't really a single trend, but a series of observations. The Color Frequency tab of the Heatmap helps visualize which colors have come in and out of use, when each color was first introduced, as well as which colors are used the most. These trends are mostly just fun to explore on the Heatmap, but an additional point of interest that is extra interesting is that we can also easily see which colors have been either discontinued or simply fallen out of use. There are a few colors that were only included in sets for a few years and then never used again, as well as colors that were used for long stretches of time but that now go entirely unused.

Continued on next page…

Analysis - Visualizations

Despite the trouble we had putting this visualization together (discussed earlier in this document), we feel that it turned out nicely. It is able to cleanly show several different trends and points of interest without being too cluttered for the amount of data points that needed to be included and it is an interesting visualization to look at and explore, so it helps add some visual interest to the webpage.

The one major limitation of the Heatmap, however, is that the size of the rectangles on the Color Frequency version. While it certainly looks cool with the small pieces, it can be a bit hard to read. If we were to redesign this one, we'd find a way to make the rectangles a bit bigger.

Conclusions

Overall, we are pleased with the results of our project. While there are certainly some limitations, especially with the Connected Scatterplot (The Squiggler), the webpage is able to effectively show some interesting trends in the Bricklink Lego Dataset.

With these visualizations we were able to show some interesting trends in the dataset and in doing so, explore the history of Legos. The webpage guides the user through some of the points of interest in the dataset with the story mode features, while also allowing users to explore some of the data themselves using tooltips and the various selection menus.

In the end, we had fun building this project and exploring Legos a bit more and we hope that you enjoyed exploring the page as well!