# Timeline of Global Protest Data by Region

By Andrew Salls, Neha Kuchipudi, Samuel Darer, Shivali Mani

CS 4804 Final Project - Interactive Data Visualization

Group Members: Andrew Salls (asalls2@wpi.edu), Neha Kuchipudi (nkuchipudi@wpi.edu), Samuel Darer (sdarer@wpi.edu), Shivali Mani (smani@wpi.edu)

# Final Project Process Book - CS4804

Table of Contents

## Initial Project Proposal

Working Title: The History of Protests

Motivation and Objectives: Discuss your motivations and reasons for choosing this project, especially any background or research interests that may have influenced your decision. Provide the primary questions you are trying to answer with your visualization. What would you like to learn and accomplish?

Over the past few years, the world has witnessed several significant protests, from the 2017 Women's March, to the 2019–20 Hong Kong protests, and even the more recent outcry against the war in Gaza. We see these events happen and we know much of the context, but these are just the tip of the proverbial iceberg. These are significant protests from our USA-centric perspective, but how many major protests have happened that we haven't known about? What do those protests say about the country and time period they happened in? While this dataset won't have all the answers, it may allow us to learn and then show others through the visualization we produce.

Data / Data Processing: From where and how are you collecting your data? If appropriate, provide a link to your data sources. Do you expect to do substantial data cleanup and transformation?

We plan on using the data on protests from the Mass Mobilization Data Project Dataverse. The project's data covers protests in 162 countries between 1990 and March 2020. For each protest event, the project records protester demands, government responses, protest location, and protester identities.

Data location: https://dataverse.harvard.edu/dataverse/MMdata

Visualization Design: How will you display your data? Provide some general ideas that you have for the visualization design.

Ideas:

- Map protests by some metric to their location.
  - Circles at protest site indicating protest size
  - Heatmap of globe by country with colors and features related to protest numbers, size, demands, government response, and so on
    - § Additional features could be embedded shapes or icons, the addition of a texture, etc.
- Flow diagram showing numbers of how the number of protesters is linked to causes, identities, regions, etc. (Like the Titanic survivors' visualization, we saw in class)
- Timeline/stream diagram derivative showing timeline of protests starting and ending with additional details based on color, texture, shade, etc.
  - This is a stretch, but possibly linked to an interactive globe visualization. Mousing over a protest and see where on the globe it took place, or mouse over a country on the globe and see what protests took place there.
  - An alternative could be creating a visualization like the cybersecurity breach visualization we saw in class (vertical timeline with circles that correlate with protest size). Added features could be searching for a location or name of a protest to filter through data.

Anticipated Challenges + Anything Else: Give us some idea of the challenges you're anticipating with this project. Feel free to describe how you plan to mitigate risk if you have ideas. Also, use this section to communicate anything else you think we might need.

- There is a lot of data to address. It will likely need cleaning of some sort and that may impact what we can represent.
- Visualizing the data in an accessible way. In the dataset there are a lot of protests that have happened in the same area, which could lead to crowding or lack of ability to interpret data in a map-based visualization. Instead of a marker for every protest, bucketing each protest at a certain area into one marker might make it more accessible.
- A lot of the protests in the dataset have the location “national level” instead of a city. How do we represent these national level protests?

· There may be potential data bias. Protest data, or any political data, can be sensitive. How do we visualize and present the data in an unbiased way that is sensitive to the groups affected?

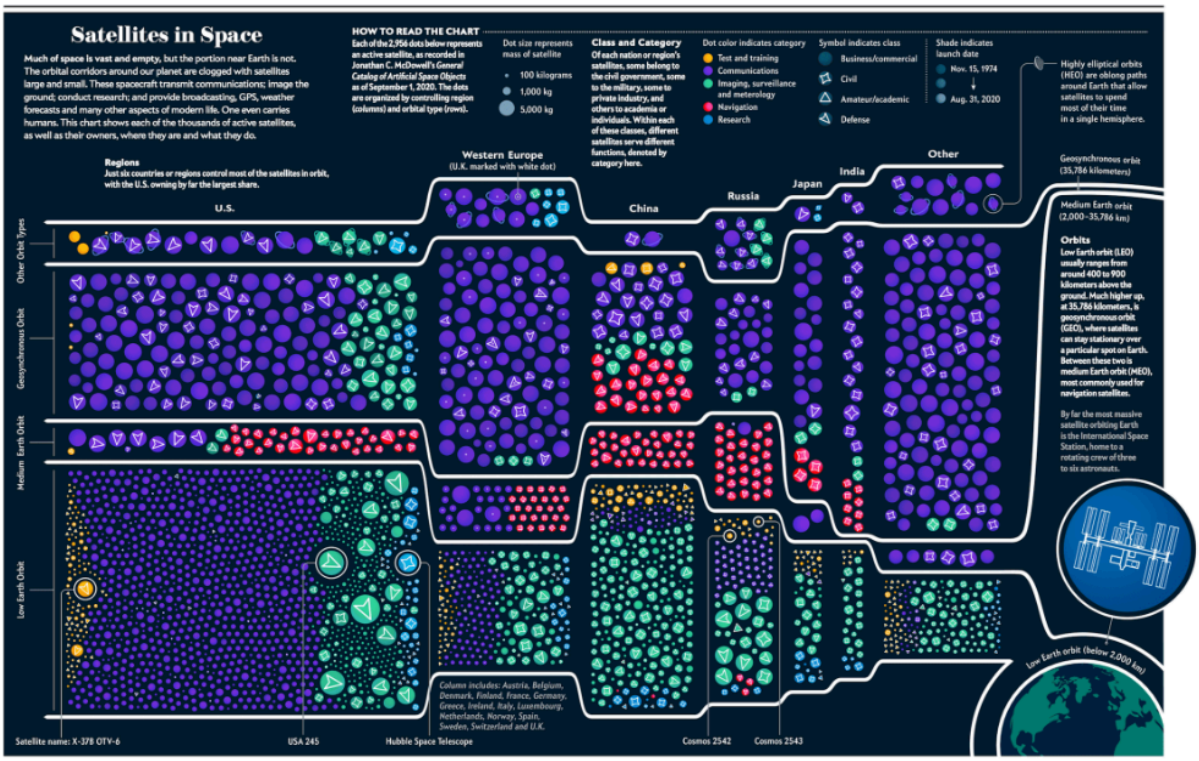
## Summary

Over the past few years, the world has witnessed several significant protests, from the 2017 Women's March, to the 2019–20 Hong Kong protests, and even the more recent outcry against the war in Gaza. We see these events happen and we know much of the context, but these are just the tip of the proverbial iceberg. These are significant protests from our USA-centric perspective, but how many major protests have happened that we haven't known about? What do those protests say about the country and time period they happened in? While this dataset won't have all the answers, it may allow us to learn and then show others through the visualization we produce.

## Research Questions

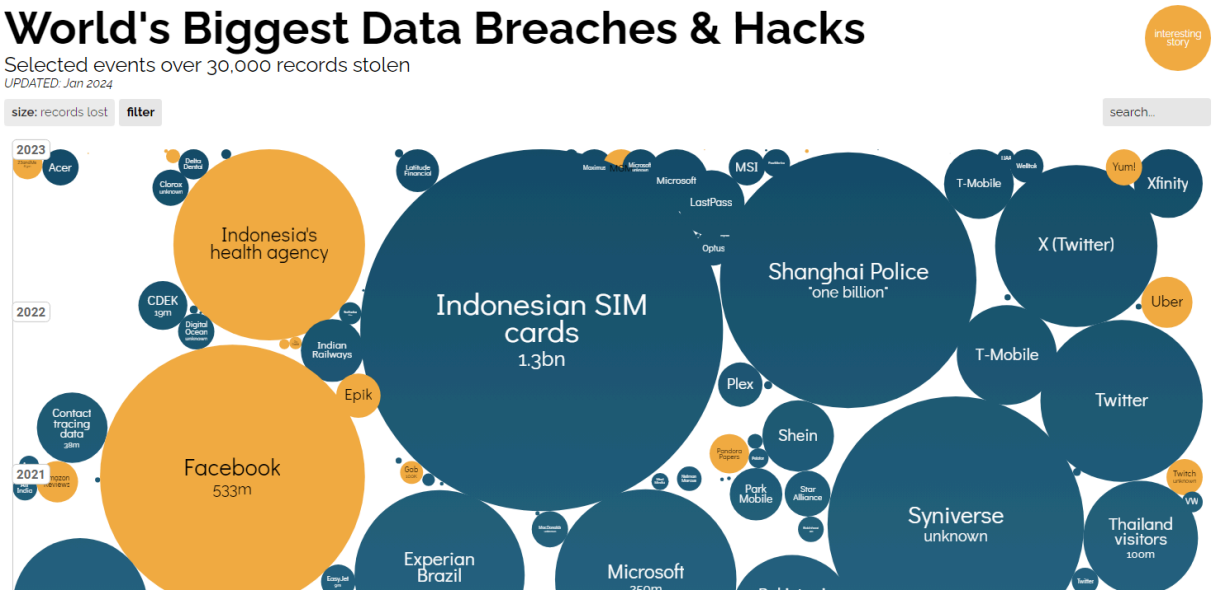
- How can we effectively show the occurrences of protests across different geographical regions with the mass amount of global protest data in the dataset?
    - How do we reduce clutter of data points?
- How do we visualize patterns of protest causes and demands?
- How do we visualize and show all of the different attributes attached to a protest?

## Inspirations




[Satellite Graphic]

This data visualization visualizes all the satellites that are still active in space. Each satellite is represented by a circle marker, with the size of the circle representing the mass of the satellite, the color indicating the category, symbol within indicating a class and the shade of the circle indicating the launch dates. The circle markers in the satellite visualization are also organized by the 7 countries or geographical regions that control the satellites and the orbit types. The organization of the markers was inspiring and relevant to the protest data due to the geographical nature of the data set. With each protest in the protest data set having so many attributes attached to it, the satellite visualization provided good inspiration on how to represent and visualize those attributes in our visualization.




[Cyber Security graphic]

This data visualization visualizes the world's biggest data breaches and hacks. Each major data breach or hack is represented by a circular marker that correlates with the number of records breached. The markers are shown on a scrollable vertical timeline. The visualization provides interactivity with hover functionality for the markers that shows additional information about each of the breaches. Filter functionality is also provided to filter breaches by the sector or method along with search functionality to search for specific breach names. The scrollable vertical timeline format provided an appealing way to present the protest data. We also want to provide interactivity with filters for the protest data by protest demand similar to this visualization.

## Data

The dataset used was sourced from https://dataverse.harvard.edu/dataverse/MMdata.

To improve the usability of the data, a number of steps were taken to clean it.

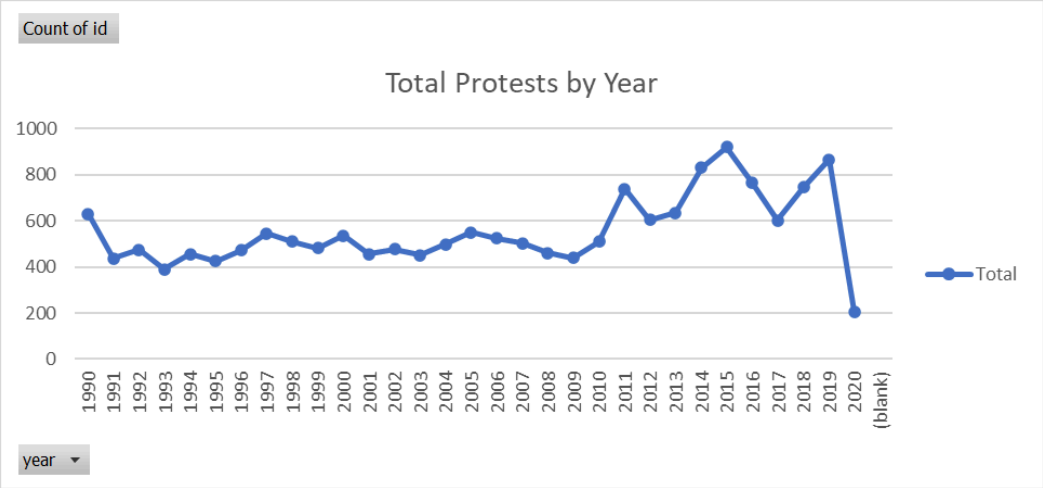
- All rows with a protest value = 0 were removed (0 indicates that a protest did not occur in a specific country in a specific year)
- Added columns startdate and enddate to have protest start and end dates in a single field. Filled values with following Excel code:
    - =DATE(year, month, day)
- Added columns startfrac and endfrac to give protest date values that were easier to plot along an axis. Month and day were converted to a decimal fraction of a year and added to the given year. The Excel code used to fill the column values is as follows:
    - =year + YEARFRAC(DATE(year, 1, 1), date, 1)
- Added column participants_helper as part of an effort to fill the blank values in the participant_category column. participants_helper removed the s from the end of values in the participant column holding strings (ex. 100s) as opposed to numbers. Excel code used is as follows:
    - =IF(ISNUMBER(participants),participants,NUMBERVALUE(LEFT(participants, LEN(participants) - 1)))
- Entries in the participant's column with ranges were given participants_category range by hand
- Any row with a participants having 1000s or a range ending in 1000s was given 2000-4999 in the participants_category column
- Any row with a participants having a range ending in 10000s was given >10000 in the participants_category column
    - In general, in rows with a participants having a ranges, the top of the range use used for determining the participants_category column value
- Remaining blank participants_category entries were filled using the following command:
    - =IF( participants_helper>10000,">10000",IF( participants_helper>4999,"5000-10000",IF( participants_helper>1999,"2000-4999",IF( participants_helper>999,"1000-1999",IF( participants_helper>99,"100-999","50-99")))))
- Added a new column participantsizeindicator based in participant_category range. The column was filled with following command:
    - =IF(participants_category=">10000", 5, IF( participants_category="5000-10000", 4, IF( participants_category="2000-4999", 3, IF( participants_category="1000-1999", 2, IF( participants_category="100-999", 1, 0)))))
- Blank values in the protestors column were filled by hand with appropriate values
- Removed columns: id, code, protest, protestnumber, startday, startmonth, startyear, endday, endmonth, endyear, participants_helper, participant_category, source, notes. This was to remove extraneous information from the data file and reduce the data file size to expedite loading in d3.

- Removed values of '.' from protesterdemands4 and stateresponse7
- Removed rows that had no value for the stateresponse1 column

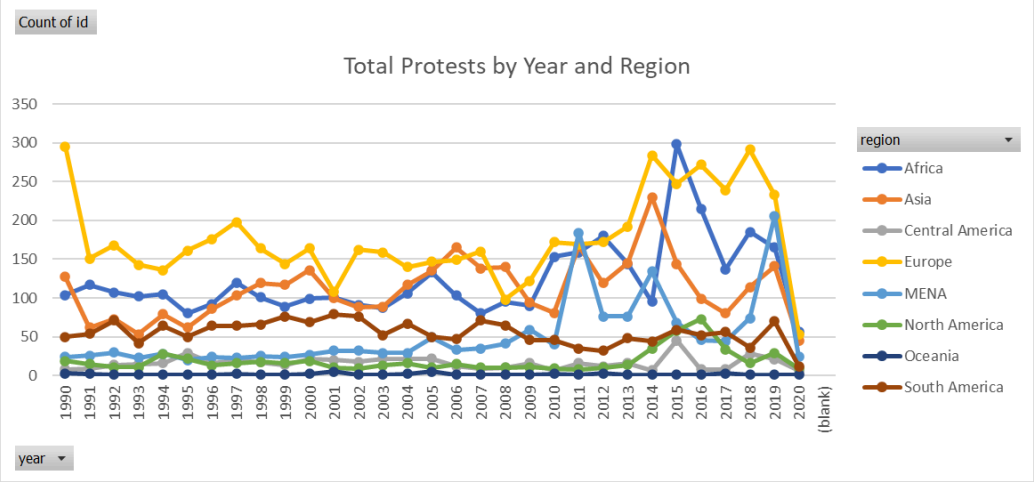
## Exploratory Data Analysis

Initially there is an existing data visualization of the dataset at https://massmobilization.github.io/. However there is no data pertaining to United State protests. Originally a data analysis was conducted in excel to understand all the features in the dataset, examples include line graphs of all protests by year, by year and region, year and country, etc. Findings include within the African countries, the large blue spike in 2015 is Kenya, the light blue one beneath it Burundi. The brown peak in 2016 is Nigeria while the brown peak in 2005 is Namibia and the pink peak in 2018 is Madagascar (Figure 3). For the countries in Asia, the darker orange peaks ~2004 in South Korea. Orange peaks 2011 and 2014 are Bangladesh. Blue in 2019 (and 2016, 2014) is China. Brown ~2012 is Kyrgyzstan (Figure 4). Given this snapshot of the data, there was a possibility to understand what affected protests becoming violent, or around what part of the year has gotten the most protests.

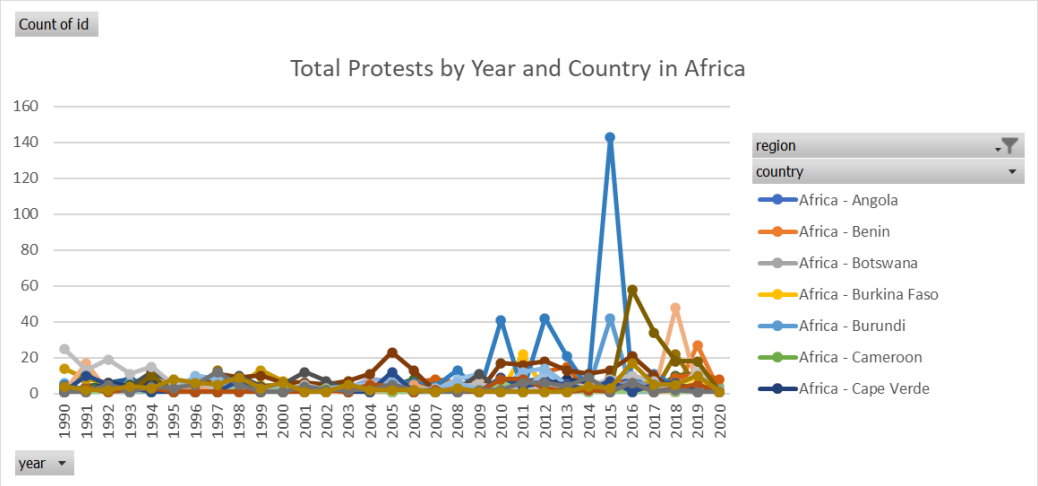
### Excel Explorations



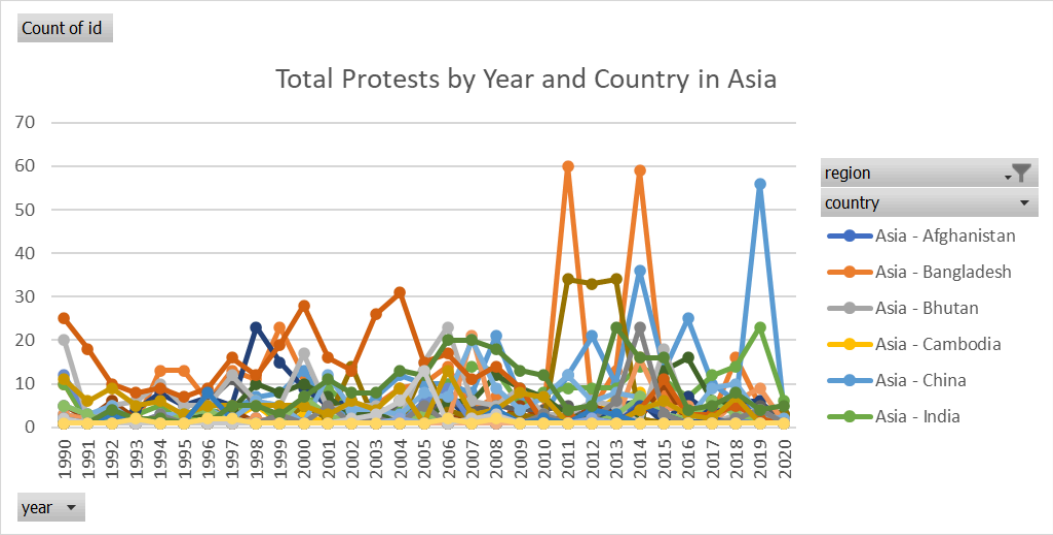

[Excel - 2/23/2024 (Figure 1)]



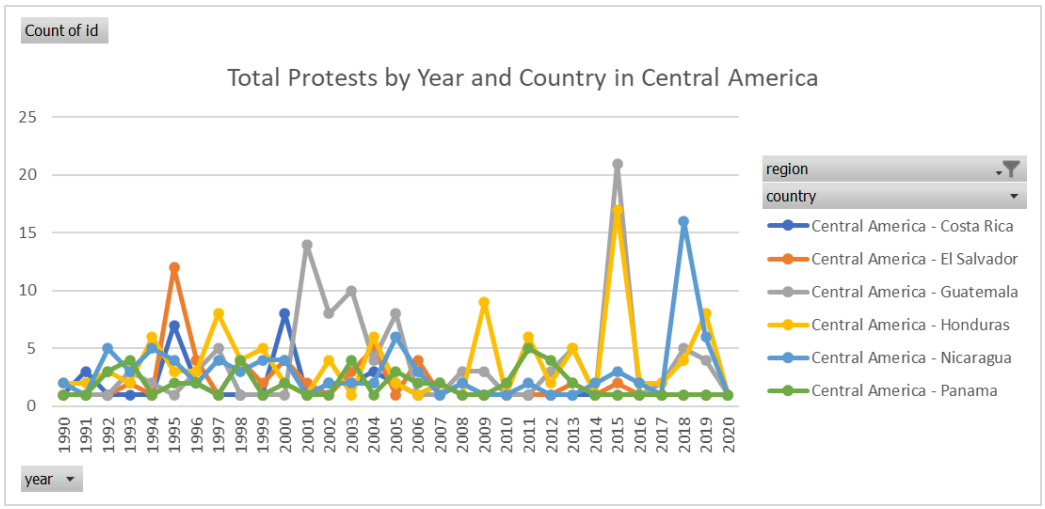
[Excel - 2/23/2024 (Figure 2)]



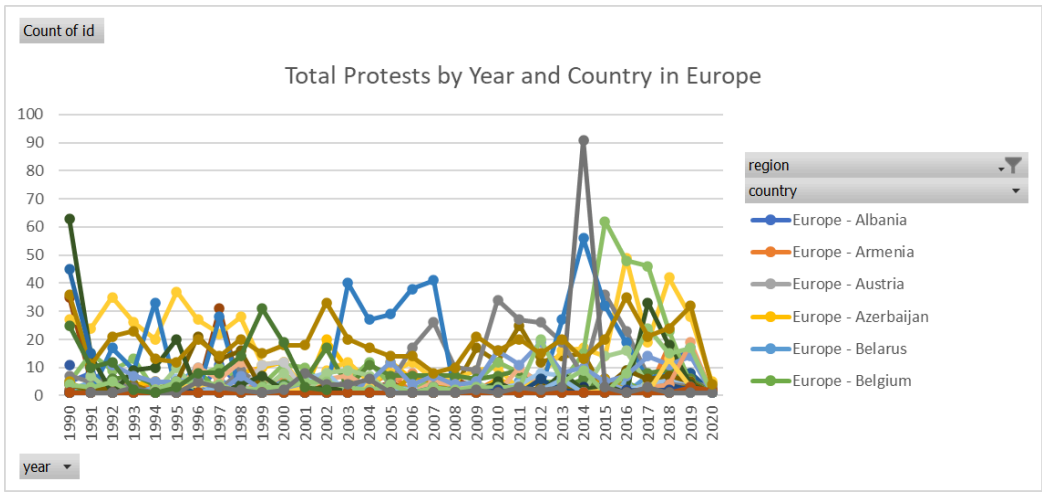

[Excel - 2/23/2024 (Figure 3)]



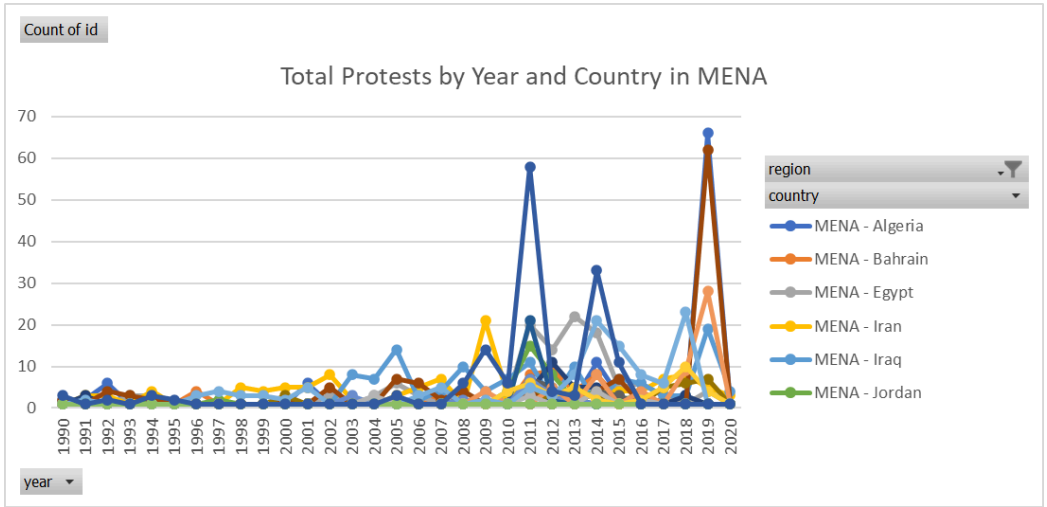

[Excel - 2/23/2024 (Figure 4)]



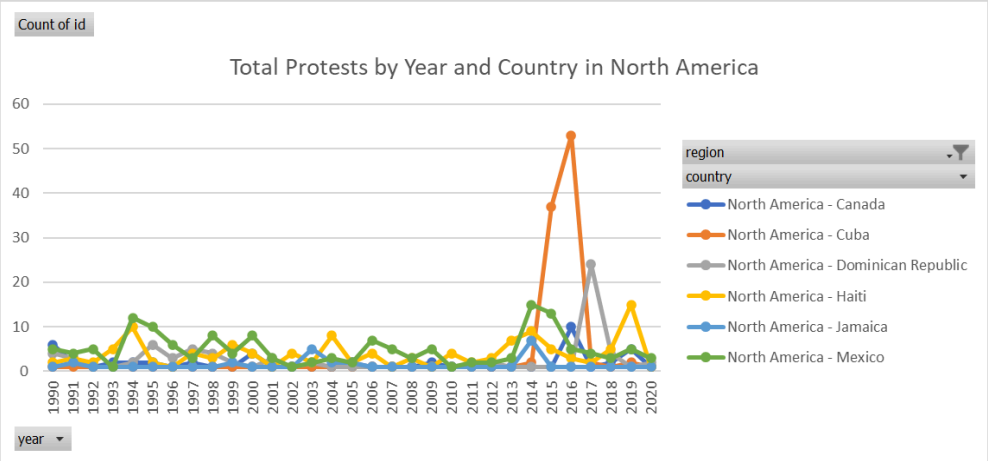

[Excel - 2/23/2024 (Figure 5)]



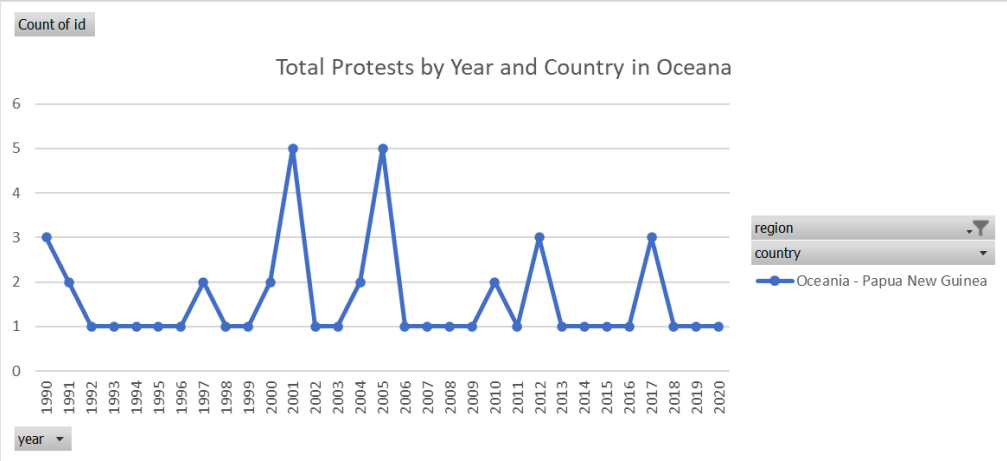

[Excel - 2/23/2024 (Figure 6)]



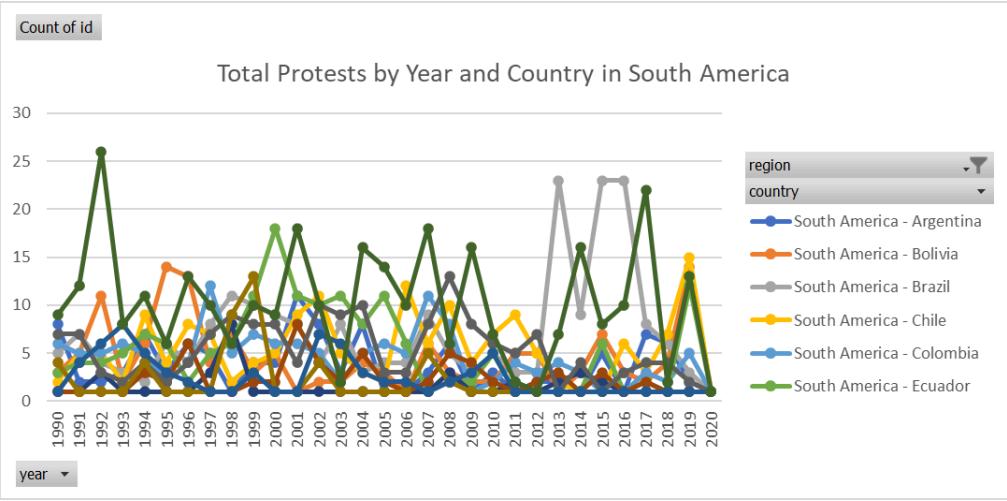

[Excel - 2/23/2024 (Figure 7)]




[Excel - 2/23/2024 (Figure 8)]




[Excel - 2/23/2024 (Figure 9)]

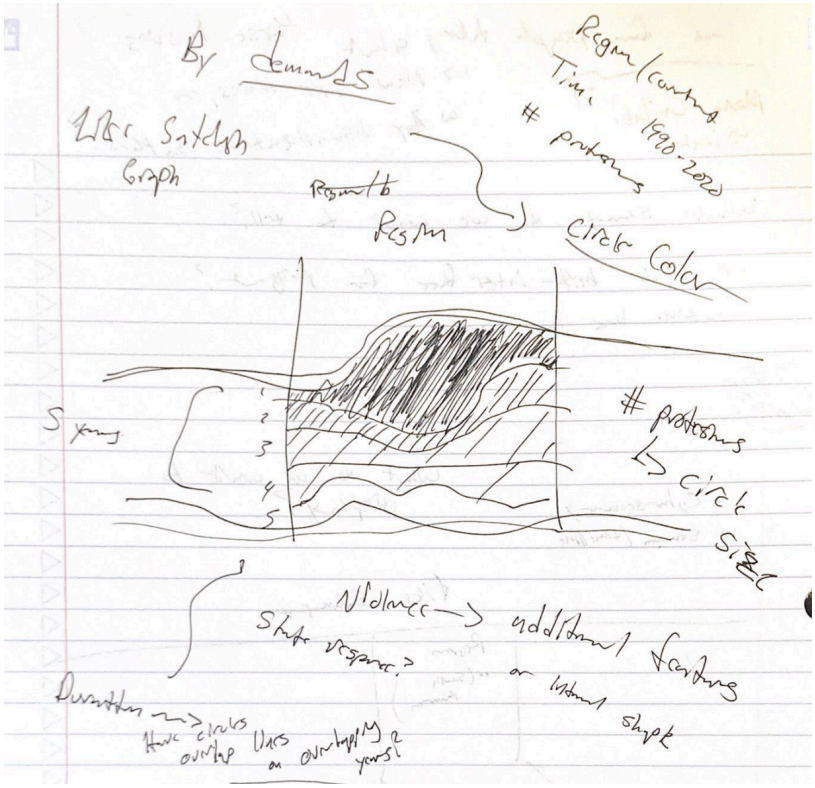



[Excel - 2/23/2024 (Figure 10)]

## Design Evolution

### Initial Sketch

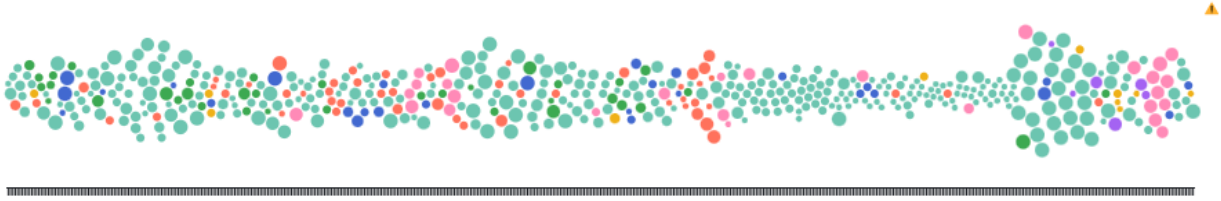
To start creating a visual that represent the data, we drew an initial design which seen below:



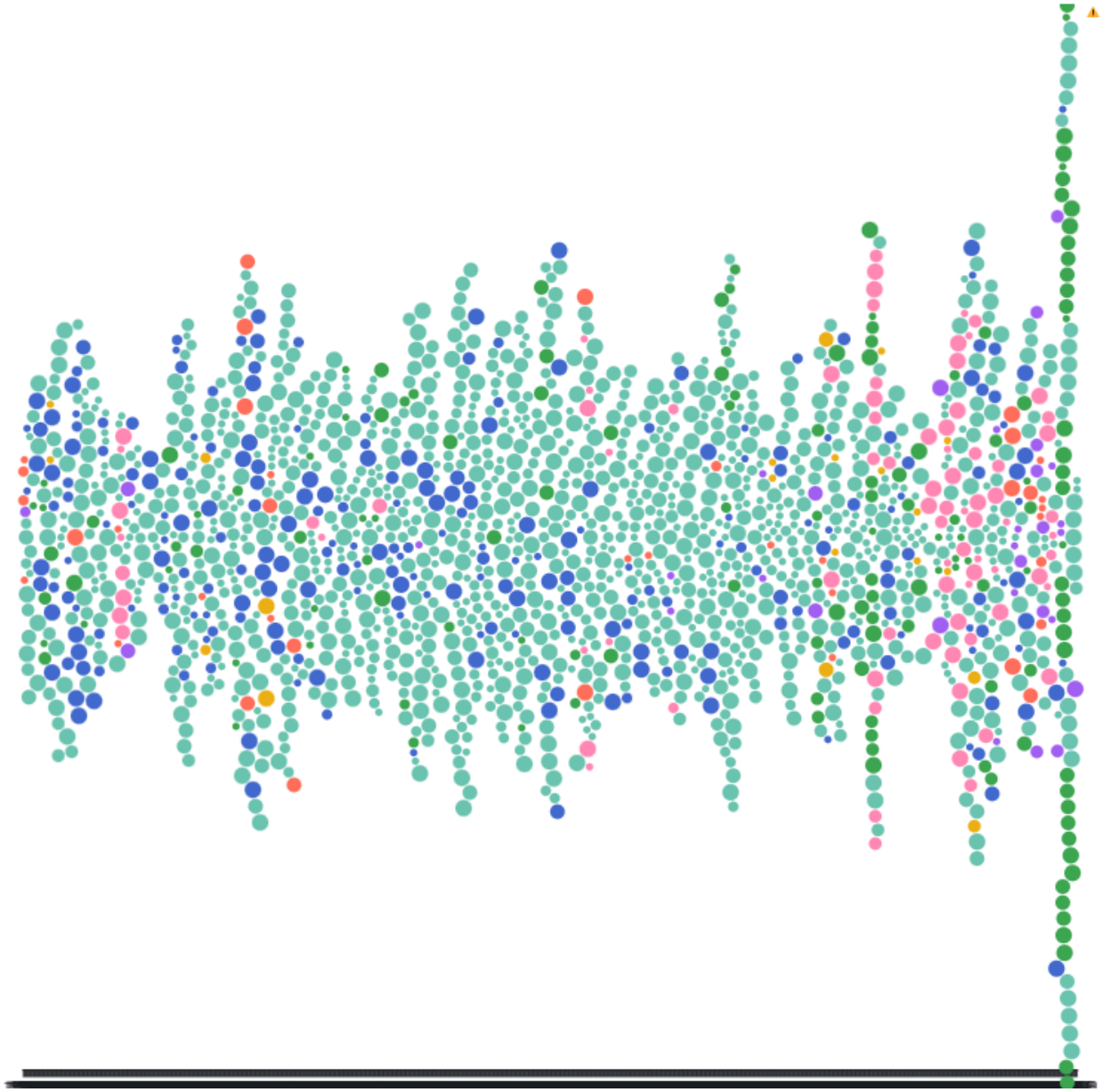
[Paper - 2/18/2024]
From the drawn visuals of a multi-faceted graph which is based on the satellite graph as mentioned in the inspiration section. The initial idea was to create a similar graph that corresponds to the protest data across the world. Since the creation of the paper version of the visual was not advanced to reference for when it needs to be coded.

### Observable Explorations

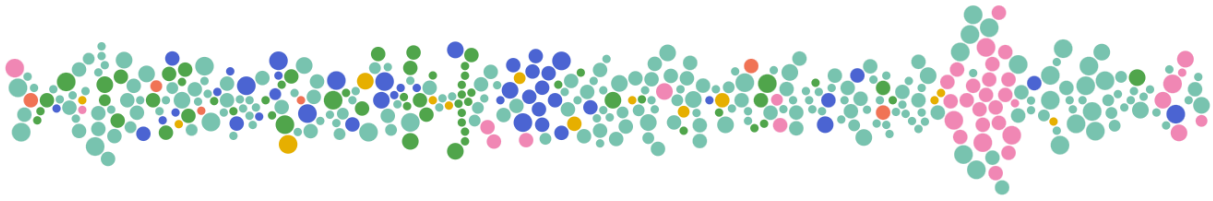
An initial method explored to present the points was using a Mirrored Beeswarm graph, using code from Observable (https://observablehq.com/@d3/beeswarm-mirrored/2).



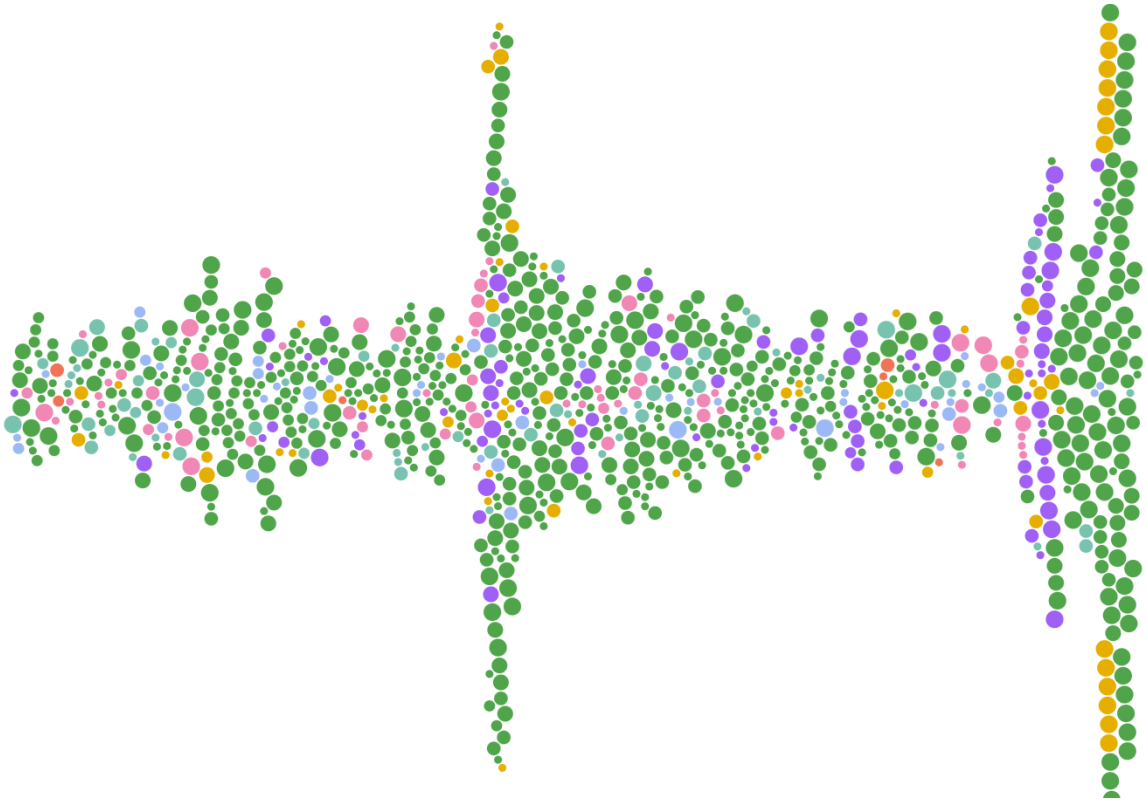
[Mirrored Beeswarm graph with region = ‘North America’ - 2/18/2024]



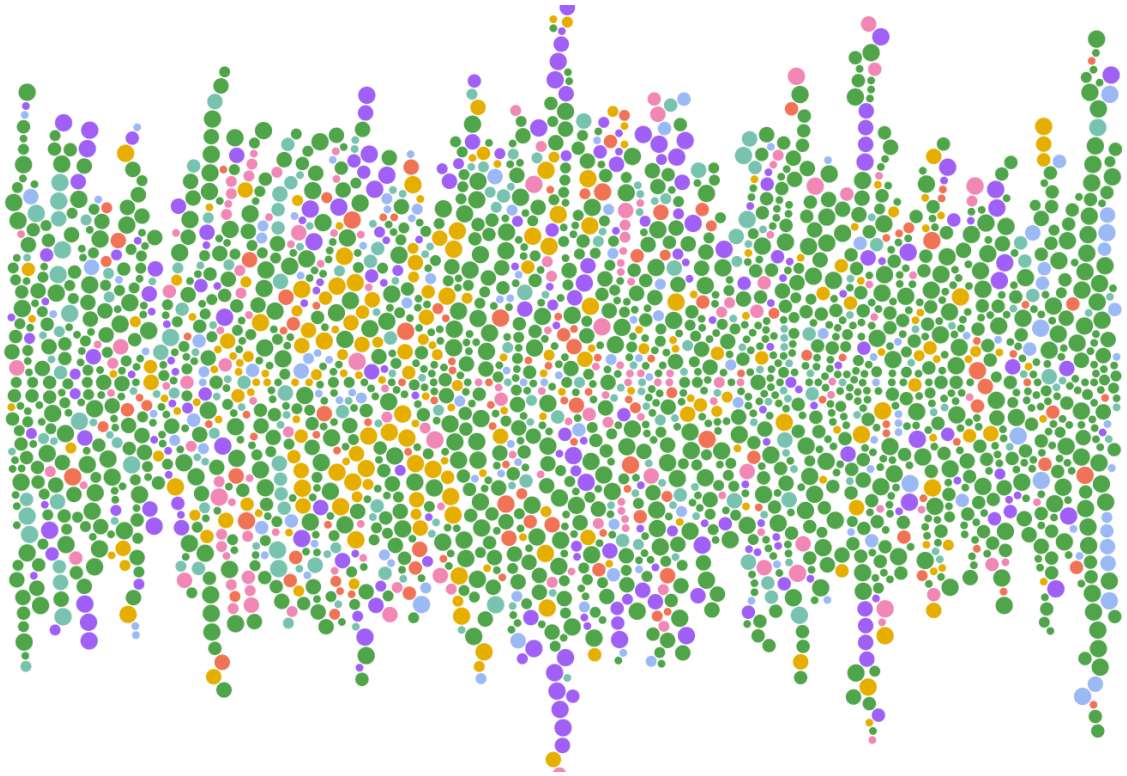
[Mirrored Beeswarm graph with region = ‘South America’ - 2/26/2024]



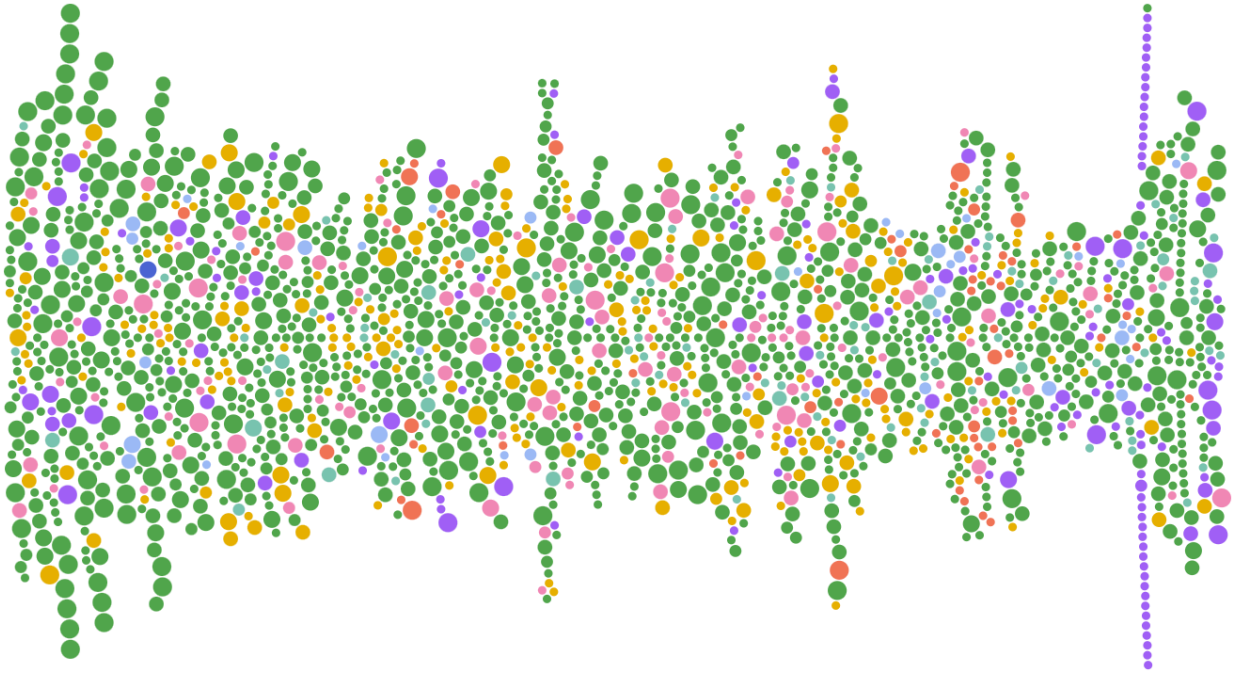
[Mirrored Beeswarm graph with region = ‘Central America’ - 2/27/2024]



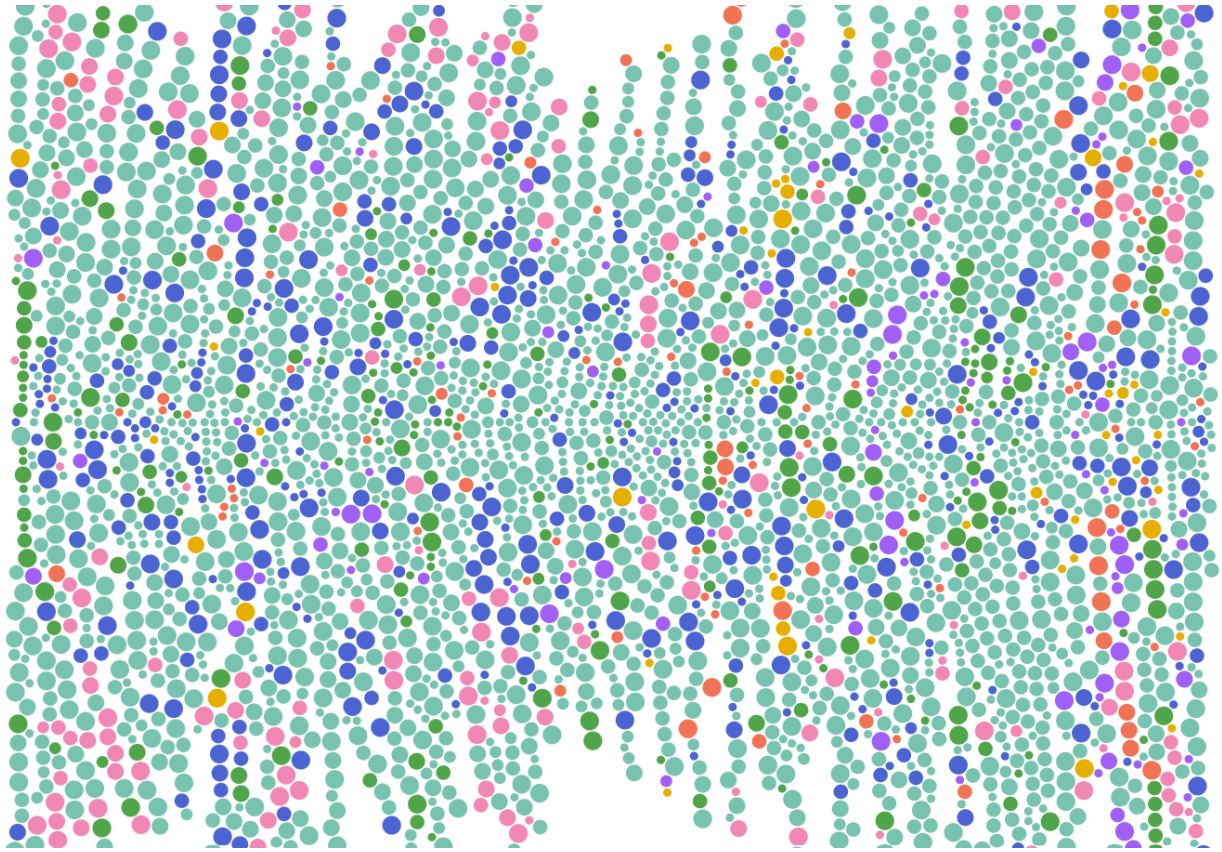
[Mirrored Beeswarm graph with region = ‘MENA’ - 2/27/2024]



[Mirrored Beeswarm graph with region = ‘Asia’ - 2/27/2024]



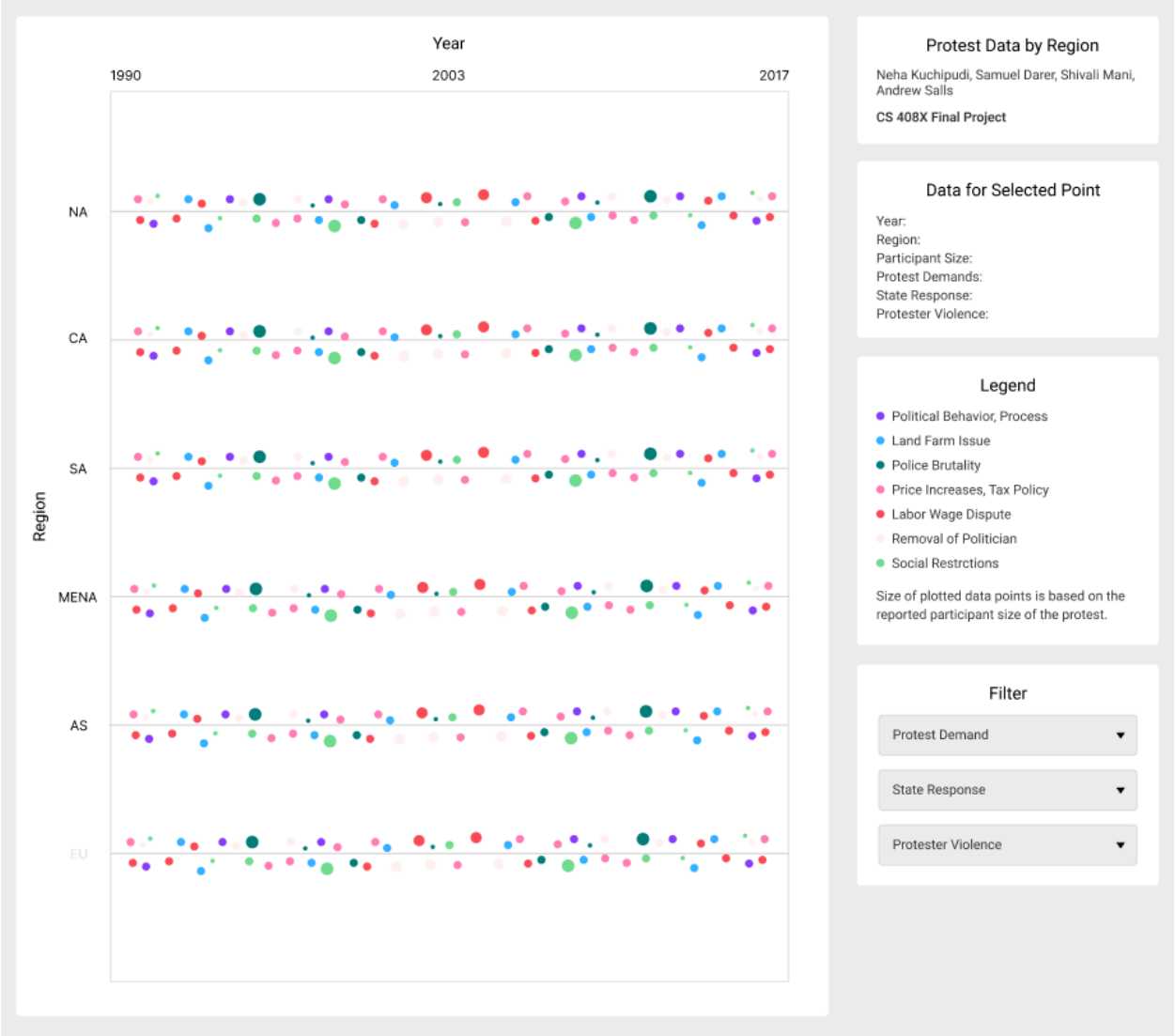
[Mirrored Beeswarm graph with region = ‘Africa’ - 2/27/2024]



[Mirrored Beeswarm graph with region = ‘Europe’ - 2/27/2024]

## Figma Mockup

The next step we took was creating a mockup layout on Figma.



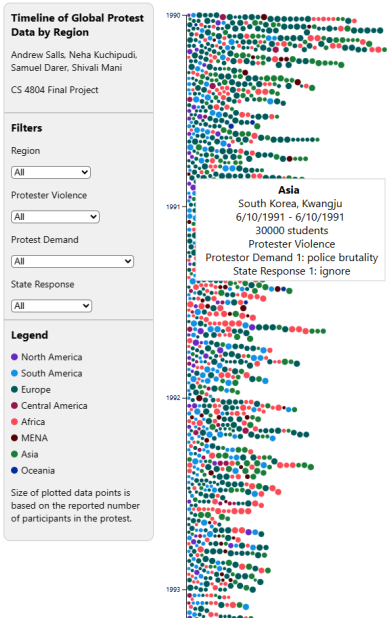

[Figma - 2/26/2024]
We utilized Figma to create a mockup of the final product of the visualization. The mockup aided us in determining a proper layout for the different components of the visualization before attempting to actually implement it. This mockup includes a section for the actual visualization, with a bee swarm graph for every geographic region in the data. The mockup also includes a section for visualization details, data for a selected point, a legend and filters for different attributes of the data.

## Technical Implementation

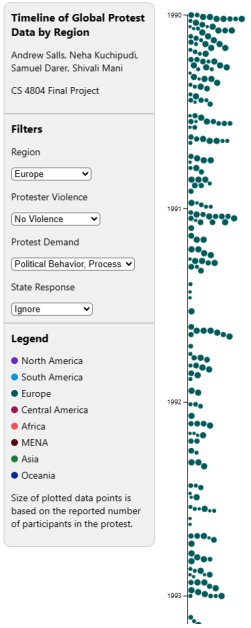
Both Observable (https://observablehq.com/explore) and ChatGPT 3.5 (https://chat.openai.com/) were used to expedite the process of prototyping and creating visualization code.

The technical implementation of our visualization closely resembles our Figma mockup. The visualization features the main graph, a vertically scrollable timeline from the year 1990 to 2020.

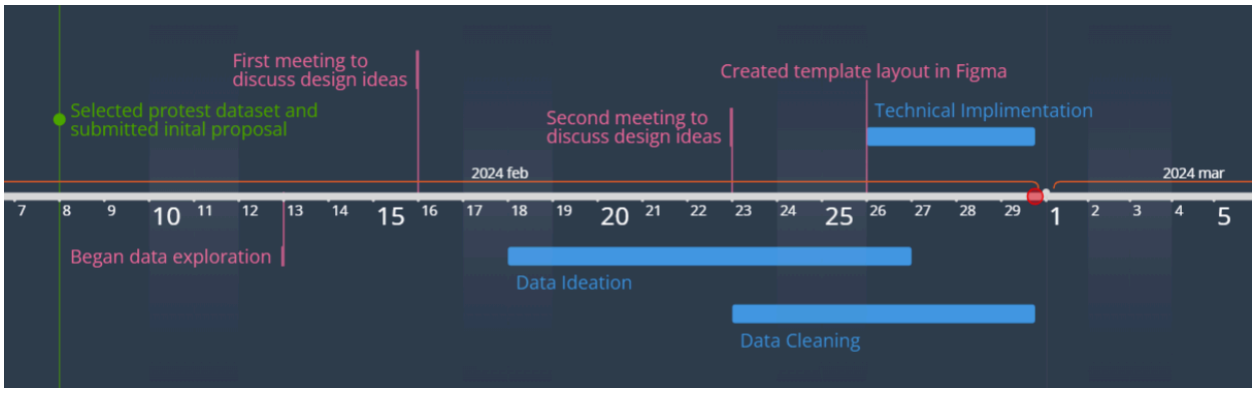

Each protest is represented by a circular marker with a color that correlates with the region the protest took place. Users can hover over a marker to view more details about the protest in a tooltip.




The visualization also has a side panel that scrolls with the timeline. The panel contains some information about the project, interactive dropdown filters to filter the graph by key data attributes (region, protester violence, protest demand and state response), and a legend.



## Timeline




Created using [Time Graphics]

## Discussion

Through working with the data to create this visualization, we did learn some interesting things. We were surprised to see the distribution of protests across the different global regions, with Europe having the most number of protests by far (4991 protests after data filtering), followed by Africa (3177 protests), Asia (3118 protests), South America (1652 protests), the Middle East and North Africa (MENA) (1256 protests), North America (527 protests), Central America (442 protests), then Oceana (38 protests). We suspect there is some amount of reporting bias in this given the relative distribution of the world population across those regions, particularly between Europe, Aisa, and Africa. In addition, protests in the United States of America were not included in the dataset. This is perhaps not surprising upon seeing that the Mass Mobilization project was sponsored by the Political Instability Task Force (PITF), which is funded by the U.S. Central Intelligence Agency.

As for our research questions, ultimately we did not manage to implement all the ideas we had with regards to visualizing the data or answer all the questions we hoped to. The large number of data points and our relative inexperience with the data visualization software meant that we did not manage to implement custom markers to show as many aspects of the data as we hoped for. In addition, revealing patterns in the data with regards to demands beyond beyond

Again due to the sheer amount of data, the graph when initially implemented was a lot wider than expected. Because of this, we were unable to divide the protest data into separate graphs based on the region in a legible way as originally proposed and designed. To circumvent this obstacle and still visualize the regions of the protests on the graph, we color coded the markers based on the regions.

We did manage to depict all the data for protests around the globe in a somewhat digestible fashion, with filters to allow for closer inquiry about various facets of the data.

There are a number of ways the visualization could be improved. One idea would be to implement a custom marker to show more attributes about each of the data points up front. Another would be to split the data into separate graphs by region and or to have a way to compare different sets of data side-by-side (ex. Violent Protests vs Nonviolent Protests). In addition, there was notes data about each protest we could not include due to the inconsistent formatting of that data. It might be possible to use natural language processing, ChatGPT, or something similar to correct the spelling/syntax/etc. of that, but it was not something we succeeded in accomplishing.