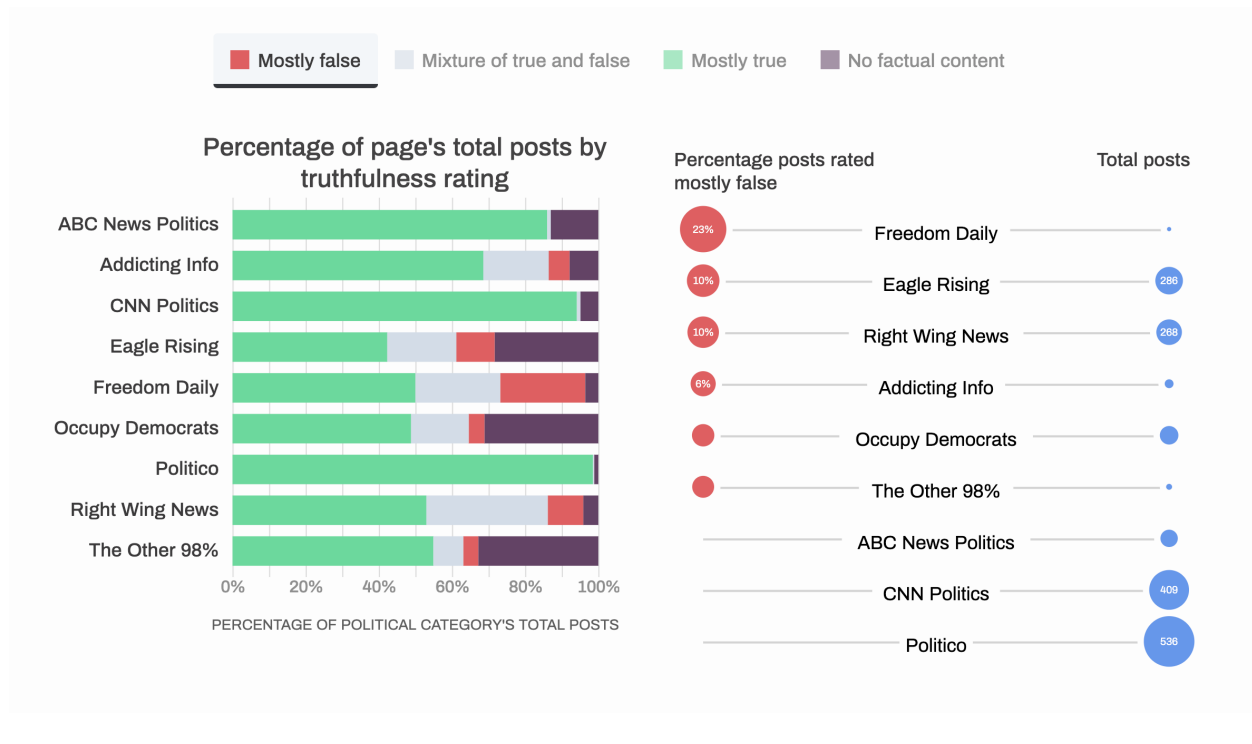




# CPSC 436V: Milestone 3

## Overview



Our group is interested in creating an interactive experience that allows the general public to explore a dataset from 9 American political pages on Facebook from September 2016, around the time of some key political moments such as the first presidential debate. The visualization will allow users to explore patterns concerning truthfulness ratings, engagement counts, and post formats from 2000+ posts from across the political spectrum. We're interested in the question "how does post truthfulness vary?" over variables such such as page, page category, number of total posts made by that page, and post format. We are also interested in whether post truthfulness is correlated with engagement, and how format and truthfulness rating affect how much impact a post has. Our

visualization, as a suite of coordinated charts, should be able to help us explore these lines of inquiry.

## Data

We will be using a [dataset that qualifies the extent of misleading information in shared by political Facebook pages](#). It analyzes about 2200 posts, which might be either photo, video, or external link, and characterizes their truthfulness (as attribute Rating, which is an ordinal variable ranging from Mostly True, Mixture of True and False, Mostly False, or No Factual Content). Mixture of True and False describes posts that combine verifiable information with speculative claims or uncorroborated claims, and No Factual Content refers to posts that are opinion, satire, or otherwise do not make any claims that are factually substantiated. These ratings were produced by raters from BuzzFeed News.

The dataset also includes information about the engagement levels of each post, detailing the number of shares (`share_count`), reactions (`reaction_count`), and comments (`comment_count`) per post, as well as the publication date (`Date Published`). It also includes a link to the content under discussion (`Post URL`).

In addition to documenting each story, the dataset specifies which page posted the information (`Page`) and whether the source type is left, right, or mainstream (`Category`).

The bare dataset is used for the bubbles visualization, which looks at every single post in detail.

In order to effectively answer questions about how different variables might affect engagement and truthfulness rating while accounting for a page's typical engagement, we will also be deriving a new variable that normalizes the proportion of posts with a certain rating against all posts by that page. This will be used for the 100% stacked bar chart. Furthermore, we dynamically compute the total count of posts per Rating for each Page. We also compute the total number of posts per Page overall, which will be for the page ranking visualization.

Since none of our tasks require the user to discern between the counts of shares, reactions, and comments — we will create a new variable (`engagement_count`) for each item which is a sum of the former three variables. To answer questions that involve comparing categories of the political spectrum, we will also derive a

new variable using the sum of the engagement counts from all Pages with the same political category, which will be used for the grouped bar charts and the asymmetrical violin plot. For the scatterplot, we calculate other similar summaries, such as the total engagement count per Rating per Page, and the total engagement count overall. These are necessary to create the small multiples scatterplots.

## Goals and tasks

1. [explore] whether publishers of political content on Facebook are being truthful or not in their content.
2. [compare] different Facebook Pages in order to understand whether a publisher's position on the political spectrum (left, centre, right) has any relation with how truthful their content is
3. [sort] pages to understand the relationship between how much they post and how many of their posts are a certain ranking
4. [derive] how what percentage of a page's posts have each rating
5. [infer] whether the truthfulness of a Page's content has an effect on how much engagement that content receives from its audience, and whether that differs between Pages across the political spectrum, at an aggregate level
6. [look up] the above relationships for a particular page, at an individual level
7. [summarize] the relationship between post format or post rating and engagement across page categories
8. [juxtapose] two small multiples, showing the effect of truthfulness rating on engagement for two different pages.
9. [query] each item and its attributes in depth and at a fine-grained level (even allowing users to find the raw data itself)

### Abstract language

- Compare attribute composition across different items (stacked bar, bubble viz, ranking viz)
- Infer trends and discover distributions in one attribute when varied across different dimensions (asymmetrical violin)

- Rank items dynamically by some attribute of interest (rank viz)
- Browse the full dataset (bubble viz)
- Lookup items of interest and juxtapose the qualities of two different items (scatterplot)
- Identify outliers (scatterplot)
- Present trends across aggregated data (bubble viz, grouped bar, asymmetrical violin)
- Present trends that exist across all items (all, bubble viz - engagement higher for no factual content posts)

## Visualization

### Exploring post engagement



The bubble visualization allows the user to explore the distribution of posts, post engagement, and truthfulness across all posts from our dataset. The task at hand is not to understand the exact proportions of accurate versus inaccurate posts, since that's what the purpose of the stacked bar is for. Rather, this visualization

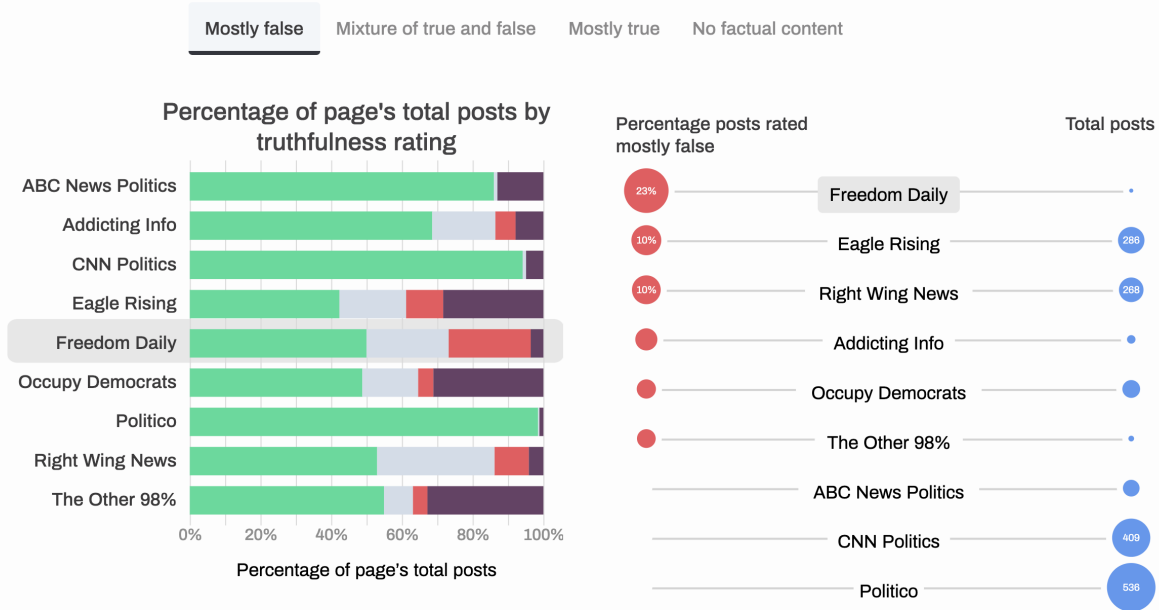
allows the user to explore individual posts and identify outliers. The default view for this visualization shows the posts grouped by their respective Pages, so the viewer has a sense of the overall composition of the truthfulness of posts when comparing across Pages. The selection interaction allows the user to group bubbles by political category, Page, or to have no grouping at all. This gives them a snapshot of which political categories or Pages have the most engagement, and what type of posts get the most engagement within those categories.

The area of each circle encodes the total engagement count for that post. Total engagement count was a derived variable calculated by summing up counts of shares, comments and reactions for each post. The domain of engagement count ranged from under a hundred to over a million. Although it's difficult to discern small differences in area between circles, due to area being a less effective perceptive channel than say, length, that is not the main purpose of this visualization. This visualization allows the user to easily spot outliers, such as posts that have gotten a lot more engagement relative to others.

Hue encodes the truthfulness rating for that post, indicating whether it was rated mostly false, true, a mixture of both, or an opinion/satirical post. We chose hue to encode the rating using hue because the cardinality of the attribute was low, and because certain hues already semantically denote the categories we working with (green for true, red for false, grey for neutral). The decision to represent "no factual content" posts using purple comes from its closeness in hue to red, without having the negative, or false connotation associated with red. Note that "no factual content" does not mean that the post was false, but that it contained content that was deemed to be entirely opinion or satire.

Upon hovering on a tooltip, the user sees the Page, the truthfulness rating, the engagement count, and the format of the post. On click, a link to the original Facebook post is opened in a new tab so the user can understand the contents of the post for themselves.

# Page accuracy visualizations



We also want to explicitly explore the interplay between post quantity and truthfulness at an overview scale, and we've chosen a new visualization, which lists each page with two point marks size-coded to represent the percentage of posts of a certain Rating made by a Page against all posts for that Page (and colour-coded to represent Rating as above). The total posts vary from about the tens to the hundreds (maxing out around 500), while percentages run the full scale from 0 to 98% or so. Both of these are quantitative, just like the raw data — evidenced by the fact that we just computed some new attributes from them. These pages are then ranked by the Rating count descending. We feel that this more saliently conveys the relationship between how much each page posts and how much of those are truthful.

For the purpose of comparing across all four Rating types, we added tabs to switch between toggle between displaying different Ratings. This also automatically reorders the elements such that the descending ranking is preserved. Therefore, which Pages have the greatest accuracy/inaccuracy/etc is always perceptually easy to understand.

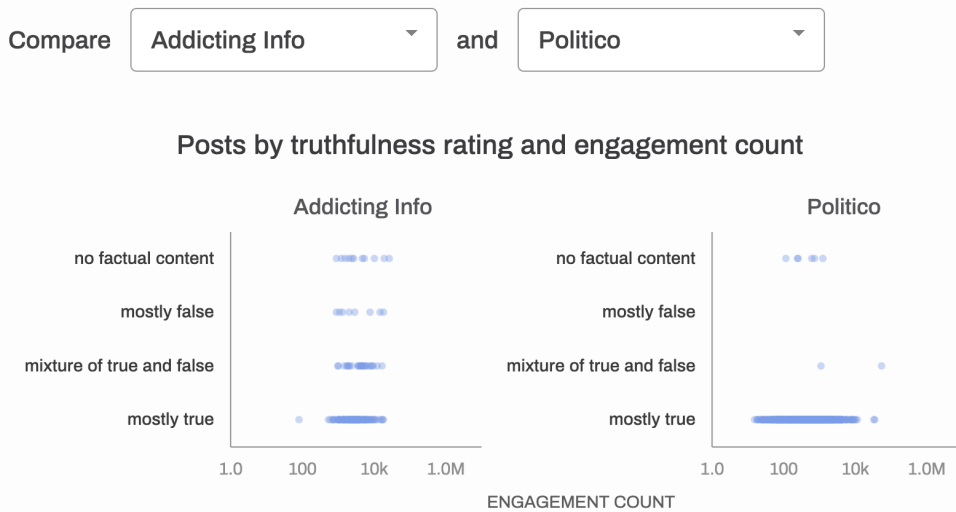
This visualization is juxtaposed against the stacked bars as a multiform detail-overview relationship. The page rankings show the percentage of Pages with a

certain Rating contrasted against total posts, which suggests how many posts a Page made as well as how truthful they are. It emphasizes how truthful a page is in the context of post volume. The 100% stacked bars drills down into a granular distribution, but does not perceptually reveal anything about totals (without hovering to produce the tooltip, it simply compares percent distributions). However, normalizing post count is indispensable to be able to accurately compare what percentage of posts made by a page are truthful. These integrate to more meaningfully reveal how truthful a Page's posts are. If we had simply made one visualization that showed both (for example, a stacked bar chart using absolute post count) it would be very difficult to perceive differences between bars that had different total lengths and difficult to accurately read values from shorter bars.

To better relate this novel ranking visualization with our stacked bar chart, we've added bidirectional highlighting at the page level when hovering over page names, so that the audience can quickly view at which ranking a page's breakdown stands, or so that they can easily explore the truthfulness breakdown of a given page. Since length is not the most perceptually effective, the bidirectional linking from stacked bar chart to ranking chart helps verify a user's impression of where a page is ranked vis-a-vis a particular truthfulness rating. Then, we added more asymmetrically-directioned highlighting (at the percentage of posts with that Rating level). One one side, from the ranking to the stacked bar chart, when hovering over a percentage circle, we highlight the corresponding segment of the stacked bar chart, which emphasizes the percentage of posts with the current Rating within a more granular breakdown of the truthfulness distribution. However, hovering over the corresponding stack segment in the stacked bar chart only highlights the page in the page ranking viz for the same reasons as the bidirectional highlighting.

Lastly, because circle area and radius are somewhat difficult to compare perceptually, we've used a square root scale for the size, and included numbers where it would be aesthetically pleasing, so that readers can ground themselves with a concrete sense of magnitude.

## **Comparing engagement and truthfulness between Pages**



The small multiples visualization lets a user compare the truthfulness rating and engagement count of posts made by any two pages, side by side. Each point on a scatter plot is a point mark that represents a single post. This visualization enables the task we set out to facilitate: at a quick glance, the user can see how much engagement a page's posts receive based on how truthful each post is, and use the dropdown menu to choose any other page to compare those trends with. Since the points are rendered at 50% opacity, it is easy to see concentrations of many points, as well as outliers.

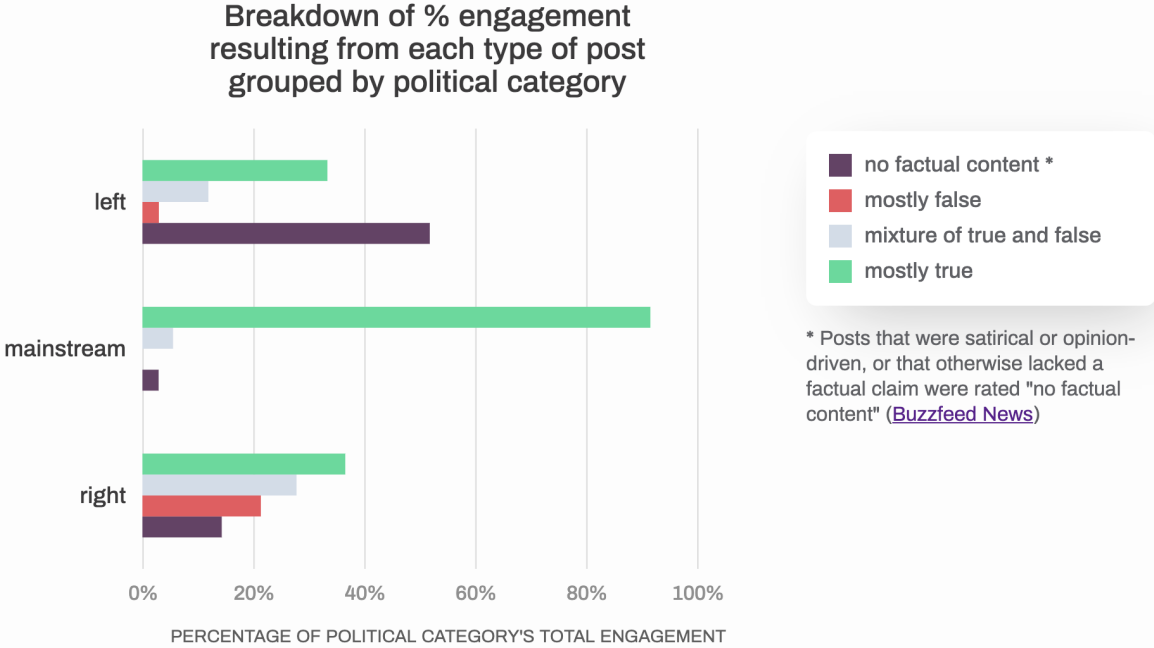
The data used in this visualization is derived from the raw data with the function `processGroupedData` in `index.js`, which groups posts by their page.

Position on horizontal scale (i.e. x-axis) encodes the total engagement for each post—how many reactions, comments and shares it received. We used a log scale for this to account for the previously mentioned wide domain of this quantitative data (less than one hundred to over one million), allowing us to fit all of our points on the same axis, but also allowing for the user to quickly spot outliers. The scale is static, i.e. we took the global maximum engagement count across all posts made by all pages, so that comparisons between pages are consistent.

Position on vertical scale (i.e. y-axis) encodes truthfulness rating. As this is an ordinal characteristic with a cardinality of four, the two extremes—"no factual content" and "mostly true"—are positioned at the top and bottom of the axis, respectively.



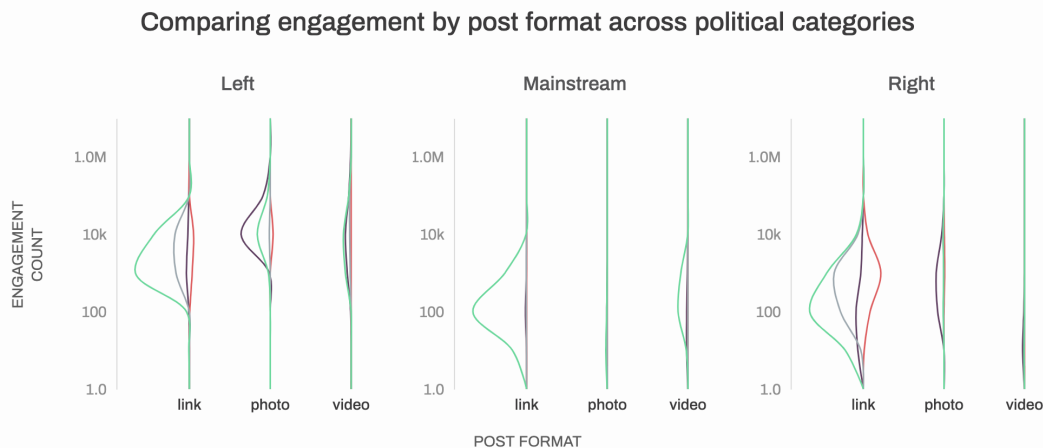
# Breakdown of engagement with posts of each ranking across political categories



The grouped bar visualization shows the distribution of engagement counts across all truthfulness rankings, for each political category. This chart reinforces certain notable phenomena shown in the bubble visualization, e.g. a little over 50% of all left-page post engagement occurs with posts that are deemed to have "no factual content," and the political category with the highest percentage of engagement resulting from "mostly false" content is the right. By aggregating the data, these kinds of takeaways are simplified to be more readily digested by the viewer.

The length of each line mark (or horizontal end position, given that the start positions are the same) encodes the percentage of total engagement attributed to posts of a given truthfulness rating. The hue of each line encodes the truthfulness rating. The position on the y-axis for each group of lines encodes which political category that group belongs to. Encoding the percentage using length — a highly effective channel for perceiving differences — allows the viewer to easily compare percentages within the same political category, and across political categories.

## Understanding engagement trends across rating and post type for each page category



Originally we were planning on using small multiple scatterplots to visualize the three dimensions of post format, rating, and page category (left, mainstream, or right) against engagement; however, the excess of points aggregated, combined with the limited screen space we had to show all three scatterplots side by side meant that the colours representing rating became muddled. It became difficult to meaningfully distinguish the colours of the points. We decided to create a visualization that would still capture the shape of engagement trends, but summarized. To this end, we used an asymmetrical violin plot, with each rating split into its own half-violin. The width (i.e. magnitude) of the violin represents a count of how many posts achieved a certain engagement count. We preserve the ability to compare engagement counts across different dimensions by aligning them on a common scale. While the grouped bar provides a general breakdown and overview of post engagement, the violin plot also uncovers distributions of how many posts got how much engagement. Is it that there are a lot of posts getting a constant number of engagement? Maybe a few posts getting a huge spike? Do some formats with some categories really succeed to capture interest?

There's a tiny amount of interactivity, just to help the viewer read some of the lines that are more occluded, if they're interested. This helps compare one rating across formats and page categories. It is admittedly subtle; however, we believe it's justified by the assertion that the interesting data really arises when there are dramatic trends. The salient perceptual information is evident at a glance.

This chart does reveal some interesting trends that might be similar or dissimilar across page categories -- for example, there are far fewer videos posted in general, and photos that contain no factual content are more common than any other kind. Also, for left and right Pages, posts with no factual content, though rarer, tended to get higher engagement. In fact, more politically partisan pages tend to get less engagement with mostly true posts, whereas mainstream links achieve an almost perfect Gaussian distribution (on a log scale) of engagement, with a lot of links hitting an engagement count of around 100. These are just examples of what inferences can be read out from the violin plot.

## Reflection

### 1. **Describe how your project has developed from your initial proposal, through your first submission, to your final product.**

We fleshed out our visualizations further after our initial proposal, expanding in detail on the information that would be shown in tooltips, how linking views would work, and the interactions involved. We also fleshed out the transitions for visualizations that benefited the most from them, such as the Page ranking viz.

At the outset, we had designed a number of component pieces, but did not know exactly how to suggest the story. Once we had all the charts in place, we were able to play around to see what order and what wording would be the most impactful. We leveraged the layout of our website to create a cohesive narrative, which we couldn't piece together until we actually saw the data visualized.

As mentioned above, we scrapped the small multiples scatterplot after realizing that the idiom wasn't appropriate given the large domain of the data and our limited screen space. We explore alternatives and came up with the violin chart which answered the same questions we were initially hoping to with the small multiples scatterplots.

### 2. **How have your visualization goals changed?**

Overall, we retained the original goals of visualizing truthfulness and engagement across many different dimensions. Some of our changes relate to being able to explore those questions more fully. We decided to pivot away from our novel square glyph chart, as it redundantly answered the same

questions as the bubble chart and to some extent, the stacked bar; instead, we replaced it with a chart that gave a sense of how much each page was posting, and how many of those were truthful. Directly visualizing the total number of posts was not something we had initially targeted, but doing so deepens an explorer's understanding of the dataset.

**3. How have your technical goals changed?**

Our technical goals have mostly remained the same since our initial proposal with the same levels and venues of interaction. Once we realized it was possible to show every individual data point in one view for our bubble visualization, we allowed that for our final implementation.

**4. How realistic was your original proposal in terms of what is technically possible in D3?**

Our original proposal was very realistic. We achieved all that we set out to do, including our stretch goals. We were pleasantly surprised to find that D3 provides essentially every method we could possibly fathom needing for our purposes. The `select` and `selectAll` functions, as well as `on`, made event handling and DOM manipulation very easy.

**5. Was there anything you wanted to implement that you ultimately couldn't figure out how to do? If so, then what workarounds did you employ, or did you abandon your original idea?**

Making our visualizations "responsive" to some extent was our biggest obstacle. A workaround we employed in order to make the rendering of the chart titles somewhat responsive was to render them using HTML text rather than SVG, as was shown during our D3 tutorials. This mitigated the problem of SVG being unable to wrap the text appropriately given the screen size. As for the visualizations themselves, we abandoned the idea of making it mobile responsive for this project.

**6. If you were to make the project again from scratch (or any other interactive visualization), what would you do differently?**

I would use CSS variables or SCSS to create variables for all the different styling that was applied to the axes, ticks, labels, and typography for consistency and ease of reuse. When we started combining our visualizations, the amount of different styling we had become difficult to manage and keep track of.

# Team Assessment

Sherry:

- Implemented bubble viz
- Visual & interaction design for website/visualizations
- Implement shared tooltip component
- Implement shared colour legend component, and tab UI
- CSS janitor - make typography and colours consistent across visualizations

Amy:

- Implemented page ranking visualization and interactions
- Added bidirectional and unidirectional highlighting between stacked bars and page ranking
- Implemented asymmetrical violin chart and interactions

Luce:

- Implemented stacked bar viz
- Added stacked bar sub-stack on-hover tooltips
- Implemented grouped bar viz
- Implemented page comparison (small multiples scatterplot) viz with page dropdown selects