Brendan J. Herger
Data Visualization
May 18, 2015

# Final Project Report

## /r/AskGayBros and you!

**Purpose:** Explore the **/r/AskGayBros** subreddit

**Data:** All user submissions (and metadata) for a 1 year period

**Tools:** Python, Pandas, LDA, D3, dc.js

**Summary:** For my final project, I've written code to download all submissions for a one year period, normalize this data, extract topics, and output it to csv.. I've also written an interactive web dashboard to explore this dataset.

# Interactivity

All charts (except for the scatterplot matrix) act as filters for all other charts, allowing the user to subset based on a number of variables. For example, clicking on the "finding a guy" topic in the Topics plot will re-render all other interactive charts to include only posts from the "finding a guy" topic.

Additionally, the topics plot provides hover over text describing the Topics, and the Day of Week, Topic Name and Adults Only charts have hover overs which communicate the number of posts in each level.

Moreover, all charts scale themselves based on the size of their bootstrap container; though implementing this was considerably more difficult than anticipated, it allows Bootstrap to handle the layout (which it access at), and the graphics to respond. This also reduces my control over plots; if a user chooses to resize the window to a weird aspect ratio, the graphics will resize accordingly. This can lead to minor issues with axis labels and squashed plots.

# Techniques

**Topics Graphic**

I've chosen to use a bubble chart to maximize the amount of information being communicated. While this leads to a large amount of whitespace, it highlights similarities and dissimilarities of the 5 topics within the primary variables of interest: number of words, number of up votes, and number of posts. Additionally, I've used color to emphasize the distinction between topics, and committed gridlines to make this graphic more fluid and approachable.

There is little lie factor associated with this graphic, as the circles are radially scaled. The data density is rather low, but this added whitespace keeps the graphic from appearing cluttered. Additionally, the data to ink ratio is relatively high. Except for removing more axis labels, there is little that can be removed from this graphic without reducing its impact.

**Length of Posts Graphic**

This histogram highlights that the vast majority of posts are pretty short, but there are a few outliers which are 2,000 words or more.

There is little lie factor associated with this graphic, as data is displayed as rectangles of uniform width. Moreover, the data density is somewhat low, but this emphasizes the shape of the distribution by complementing it with whitespace. Additionally, the data to ink ratio is relatively high; while some axis labels could be removed from the y-axis, I feel that the axis labels are necessary to provide context and guide brushing.

**Number of submissions over time Graphic**

This area chart communicates the periodicity and upward trend of number of posts over time. I chose an area chart instead of a line chart to emphasize the 'mass' of posts, and that each post represents a completed action.

I've chosen to aggregate on the day level (rather than hourly or monthly) so that this graphic can play well with the day level filter, without appearing too jagged. Additionally, I chose to label the x-axis with month names (as opposed to month number or only labeling on the year level) to provide context and assist with brushing. Additionally, there is a minor issue with the x-axis label hitting tick labels; I've chosen not to address this issue as it would require considerably more time than I am willing to invest in a minor (though annoying) issue.

There is little lie factor associated with this graphic, as the daily aggregation rate is maintained throughout the graphic. Additionally, the data density is rather high; though the line encodes most information, the area below provides context. Finally, the data to ink ratio is also rather high, once again because the area is able to support the line in providing information.

**Scatterplot Matrix Grahpic**

I also wanted to provide a high level look at how submission length and popularity interacted. While it would be possible to provide plots for each of these areas, I thought a scatterplot matrix would provide a better high level summary. I've also chosen to have this chart not interact with with the other charts; though it would be interesting doing so would be somewhat difficult (due to Crossfilter's unique paradigm), and would require me to build another plot to meet this projects 'pure D3' requirement.

I've elected to create rather large subplots with opaque individual dots to emphasize that the bulk of the posts exist in a small domain, while a few outliers exist. Additionally, I've elected to use a few light gridlines to allow for brief comparisons and sanity checks.

There is some lie factor associated with this graphic. A few outliers in each of the distributions push the bulk of posts into one position; even with opacity, this over emphasizes the importance of outliers as the central areas become saturated. Moreover, the data density is rather low, similarly due to outliers. While it would be possible to remove outliers and improve both of these metrics, I felt that it was important to view the entire distribution, and communicate that there are a few odd cases. Finally, the dat to ink ratio is somewhat low. It would be possible to remove the subplot bound boxes and gridlines, but this would remove much of the context that makes this graphic useful.

**Filtering Plots Graphic**

In my mind, the Day of Week, Topic Name and Adults only plots act as one functional unit. As such I have combined them for this analysis.

I also wanted to allow the user to be able to explore trends based on a few metadata type variables. Ultimately, I chose to include these as a fixed position footer to make them constantly available, such that the user can maintain selections between sections.

I also chose to color week days and weekends along different scales, to subtly communicate the differences between these levels. I also maintained coloration between the Topic Name chart and the Topic bubble chart. Finally, I changed the label text color for the Adults Only chart, so that it would be readable against a white background.

There is little lie factor associated to these charts; within each chart information is encoded solely as length. Though bar height varies between charts, each chart encodes different kinds of information and inter-chart comparison is not intuitive. Moreover, the data density is rather high. By elastically scaling x axis of each chart, the majority of the chart is filled with information (as long as one level is not significantly larger than its peers). Finally, the data to ink ratio is similarly high. Little could be removed other than the axis line and gridlines, though these help to provide context and hierarchy.

# Feedback

The prototype I discussed in class was pretty rough, but it still helped me to communicate my intentions to my peers. Unfortunately, they seemed more preoccupied with discussing the areas where my prototype's successes than its failures, leaving me with no suggested improvements or new areas of concern.

It's nice to feel loved, but I would have preferred constructive criticism.

# Challenges

**Data Gathering**

It was much more difficult than expected to pull the submissions and their metadata from Reddit's API using PRAW. The library's documentation is quite shoddy, and I couldn't find examples of similar projects.

Additionally, finding and using a Latent Dirichlet Allocation library proved to be more difficult than expected. I ended up using Pandas for data management, NLTK for lemmatization, SKLearn for word vectorization, and LDA for modeling.

**Web Development**

I also had an interesting time building out plots with dc.js. Though designed to be a one stop shop library, dc.js's documentation is pretty awful and there is little to go off of besides their (highly optimized) examples. Ultimately, a lot of dc.js tailoring is still done with D3, and fixing small issues (such as color consistency between charts) requires convoluted approaches.

Moveover, understanding Crossfilter's paradigm required significantly more time than I had originally allocated. Though their paradigm makes sense from an engineering standpoint, it is almost incomprehensible from a data usage standpoint.

Additionally, there were a number of small html issues which kept me quite busy. A statically placed footer is possible with Bootstrap, but difficult to work out. Similarly, modifying titles to show in a custom shade of blue required meddling in css files with a try and reload approach.

# Conclusion

This has been an enjoyable foray into the world of data visualization. I have been able to explore the high level of control which HTML5 and D3 afford, while exploring an interesting dataset.

However, if I had to do this project over, I would likely have proceed with a Tableau dashboard. Though it pains me to say this, almost everything that I've painstakingly done could be achieved quite quickly with Tableau's well designed framework. I understand that D3 is an incredibly powerful tool, but in terms of making data accessible I do not think it is worth the considerable boilerplate code and effort.