Judd Antin
October 13, 2005

Information Visualization and Presentation

Assignment 2: Data Exploration

During the course of this assignment I used two pieces of data visualization software, Spotfire and Tableau, to explore the following hypotheses:

1. Candidates running for open seats will have fewer total receipts than either incumbents or challengers but more major party contributions.
2. Overall expenditures are higher in cases where there is a runoff election.
3. Major party contributions, but not individual contributions, will be larger in states with early primaries (Iowa and New Hampshire).
4. General election percentage and total receipts are positively correlated.

As it turned out, verifying these hypotheses was fairly straightforward, and did not require any particular exploration of the data with the visualization tools. Hypothesis #4, however, provided an opportunity to explore interesting and subtle patterns, as I will describe below.

Exploring these hypotheses with Spotfire and Eureka also provided some useful insights on the strengths and weaknesses of the software. These will be described at the end of this paper.

Note: For the purposes of visualizations in this paper, the data was restricted to democratic and republican candidates only.

Hypothesis 1

Candidates running for open seats will have fewer total receipts than either incumbents or challengers but more major party contributions.

Simple bar charts are enough to explore this hypothesis. They are perhaps most effective viewed side by side as in Figure 1, generated by Spotfire.

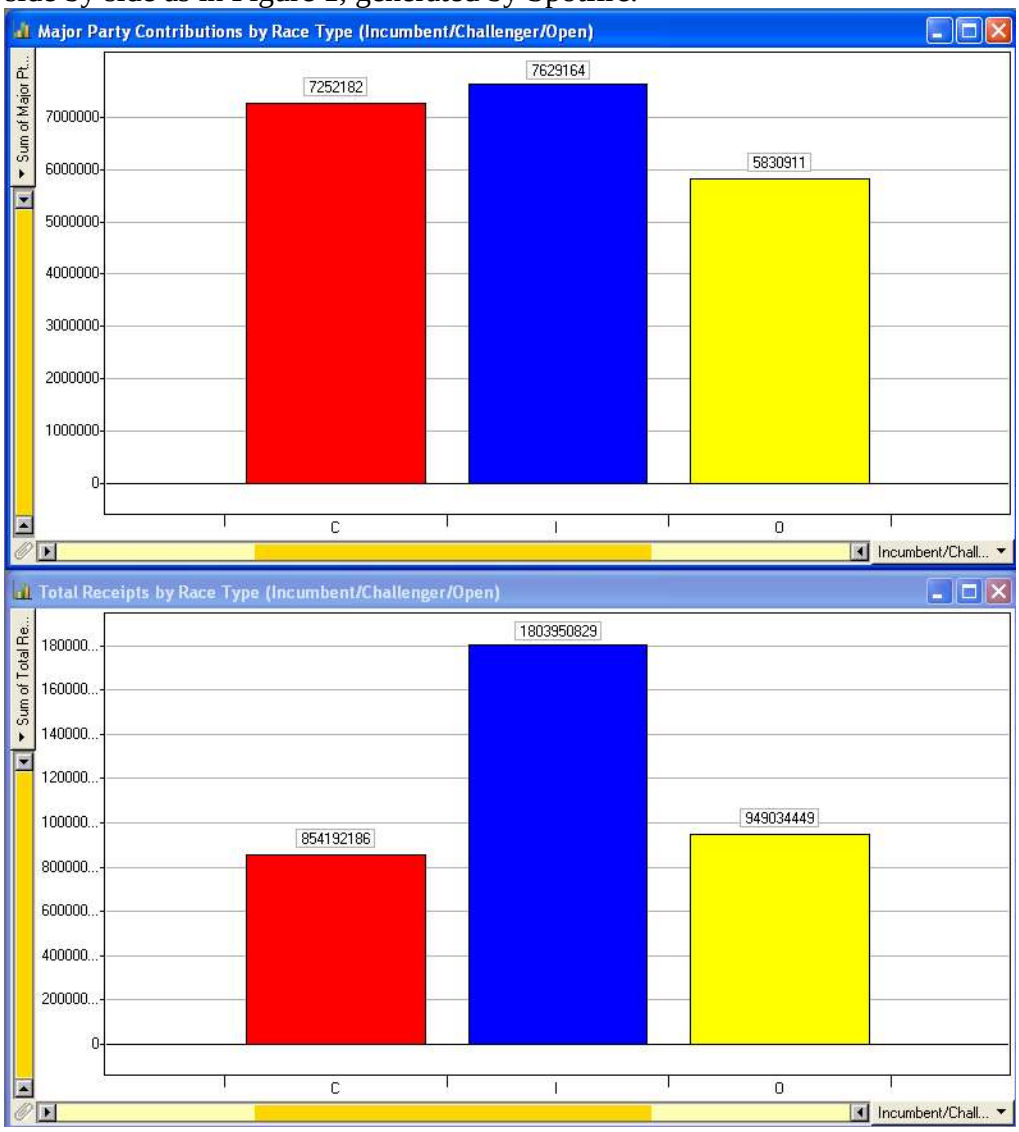

Figure 1
Total Receipts (bottom) and Major Party Contributions (top) by Race Type

The data shows that this hypothesis is flat wrong. Open seat candidates (yellow) have more total receipts than challenger (red) candidates do but far less than incumbents (blue), and fewer major party contributions than either incumbents or challengers. My original logic, that open seat candidates would be less able to generate campaign contributions through connections than their entrenched counterparts, but that the political parties would help to make up the difference, was apparently misguided.

Hypothesis 2

Overall expenditures are higher in cases where there is a runoff election.

In order to explore this hypothesis, I first created a new variable which would allow me to easily filter runoff and non-runoff elections. Exploring Spotfire, I noticed that it is possible to create new columns in the data from within the program, including columns which are generated via complex formulas and logical tests. This is an extremely convenient feature which avoids the need to alter the data externally and reload it into Spotfire. However, I ran into some trouble because Spotfire does a poor job of previewing the results of the formula. After many unsuccessful tries at constructing the correct formula, I gave up and created it quickly in Excel.

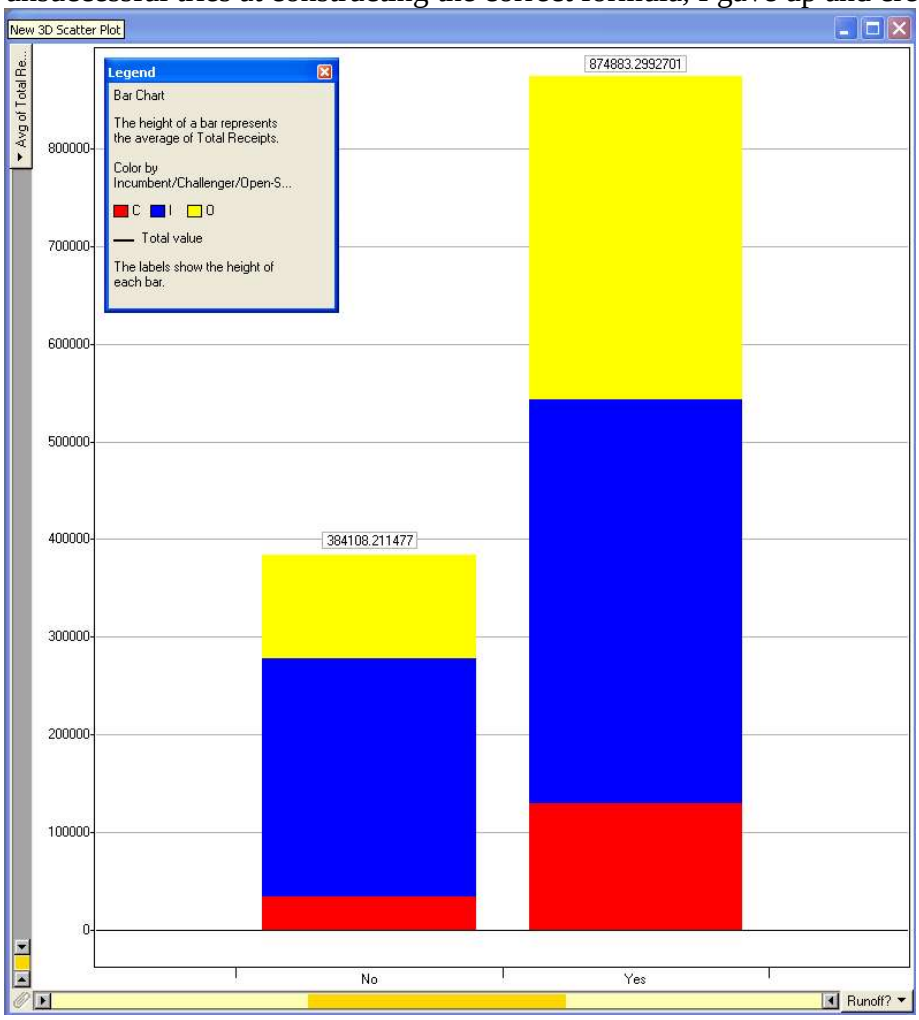

Figure 2
Average total receipts by runoff status

Figure 2 confirms hypothesis 2, and furthermore shows that it is true regardless of election type (incumbent/challenger/open). Unfortunately, Spotfire does not allow customization of the legend. As such the legend is hard to read and includes unnecessary information. In general Spotfire seems to do a poor job of allowing the user to customize a visualization's annotations.

Hypothesis 3

Both major party contributions and individual contributions will be larger in states with early primaries (Iowa and New Hampshire).

To create a visualization to test this hypothesis I once again created a simple filter variable which would separate Iowa and New Hampshire from other states. I experimented with Tableau's filter functionality for some time, trying to work within the program instead of creating a filter variable through Excel, but had no luck. I suspect Tableau has this functionality, but I couldn't work it out and I found the help files to be inadequate.

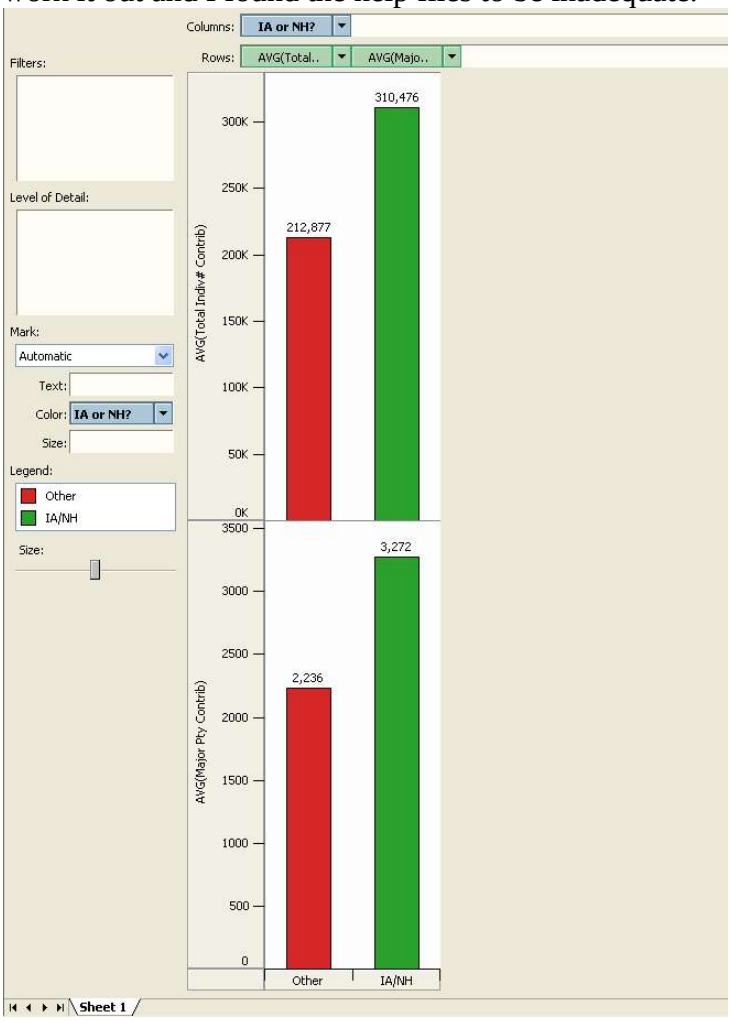

Figure 3

Average Total Individual Contributions (top) and Average Major Party Contributions (bottom) in Iowa and New Hampshire (green) and in all other states (red).

By viewing the aggregate data presented in Figure 3 we can see that hypothesis 3 is in fact true: Iowa and New Hampshire together have significantly higher average total individual contributions and major party contributions than other states do. Wondering if these numbers were as clear cut as they seemed, I took advantage of the ease with which it is possible to view small multiples in Tableau to break the data down by year.

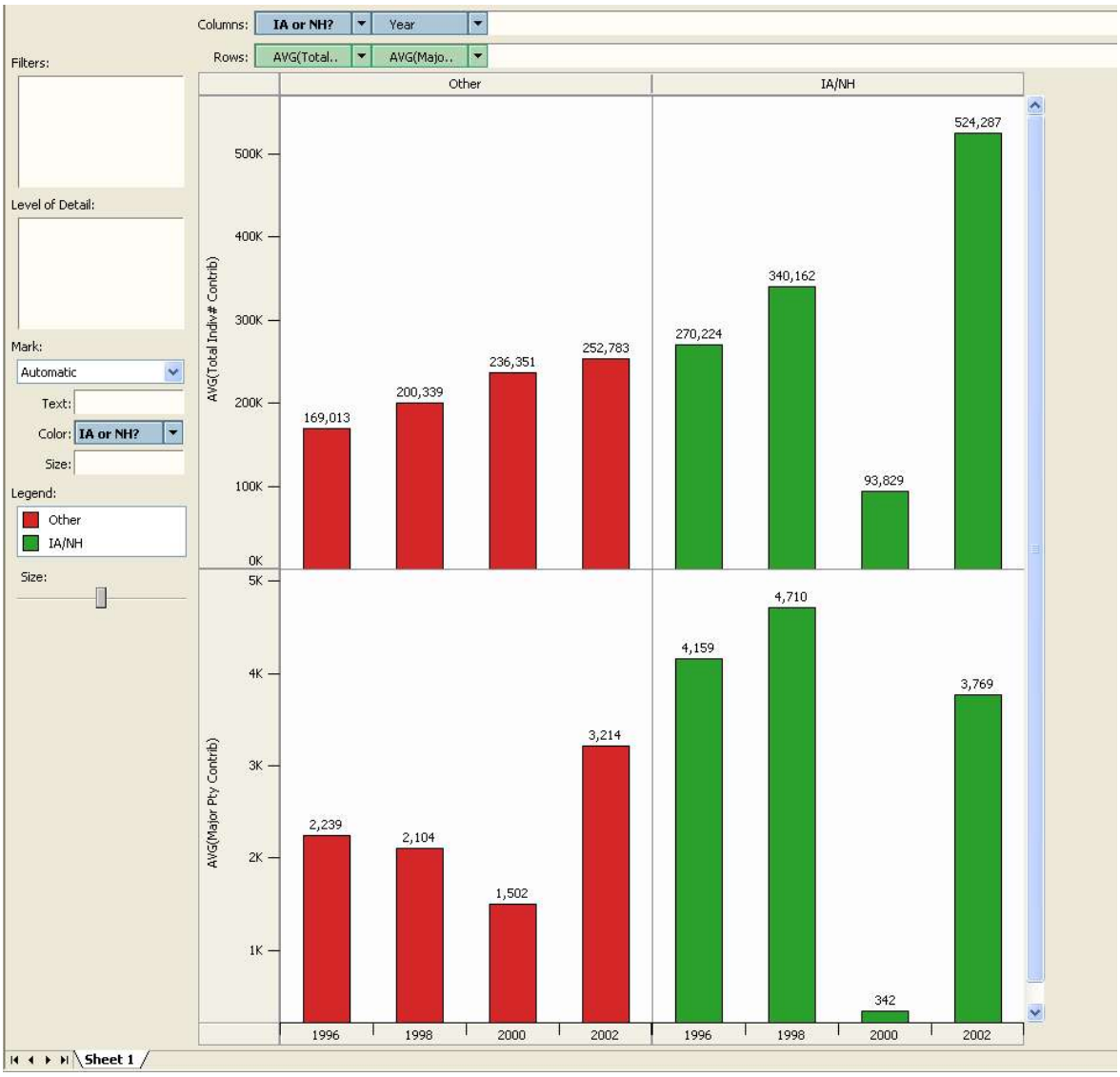

Figure 4

Average Total Individual Contributions (top) and Average Major Party Contributions (bottom) in Iowa and New Hampshire (green) and in all other states (red) by year.

Figure 4 shows that while the overall trend is legitimate, there is quite a bit more variation across years than one might expect – and in a strange place. Because the IA and NH primaries set the tone for the rest of the election season, one might expect more contributions to be centered there during presidential election years. 1996 data shows that major party and individual contributions are significantly higher on average in IA and NH than in other states. Strangely, the opposite is true in 2000, a year which I would expect funding levels to increase because of the open-seat presidential election. I am not sure how to explain this phenomenon. One possible explanation is that a much greater percentage of total contributions was directed towards the hotly contested presidential race in 2000, leaving less for the congressional races included in the FEC data. A narrow focus on the presidential election would explain in particular the drastic drop in major party contributions in 2000. However, more data would be required to confirm this hypothesis.

Hypothesis 4

General election percentage and total receipts are positively correlated.

This hypothesis provided by far the most interesting opportunity to explore the FEC data. When we use Spotfire to visualize a scatterplot where the x-axis represents total receipts and the y-axis represents general election percentage an interesting pattern emerges. (See Figure 5)

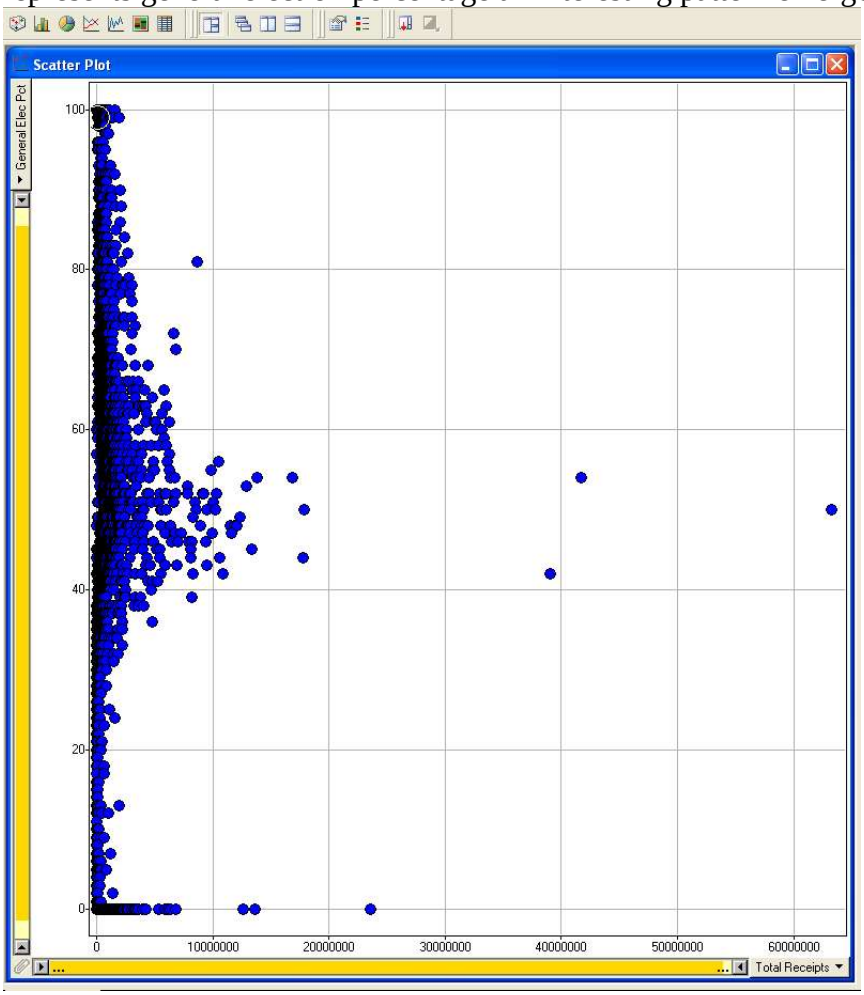

Figure 5
Total receipts (x-axis) by general election percentage (y-axis)

A few conclusions based on this visualization can be immediately drawn:

- Most candidates, regardless of the number of votes they ultimately get, spend less than 300,000. A proper visualization of this phenomenon can be found in Figure 6.
- Only 4 candidates spent more than $20,000,000 during any of the 4 elections in the data set. Each of these occurred during the 2000 election which was particularly fraught because of the end of Clinton's presidency and the battle for the White House. These contextual factors are important to consider. These 4 candidates represent interesting cases:
    1. John Corzine, Senator from New Jersey, was an independently wealthy political newcomer who essentially bought his way into the election. He spent

more than $63 million during his campaign, which was more than 10 times the average amount spent by all candidates that year. More than $60 million of that amount is classified as ‘Loans from Candidate.’

2. Hilary Clinton ran for an open Senate seat in New York on the heels of her husband’s presidency and spent more than $42 million. There was some controversy because the Clinton’s specifically took up residence in NY so that Hilary would be eligible for this seat. One might speculate that the comparatively huge amount of money spent by her campaign was an effort to combat that controversy.
3. Clinton’s opponent in that race was relative unknown Rick Lazio, who despite spending nearly as much as Clinton managed only 42% of the vote
4. Rudy Giuliani is listed as having spent nearly $25 million on his 2000 campaign despite the fact that he was not actually up for election. This could be an error in the data?

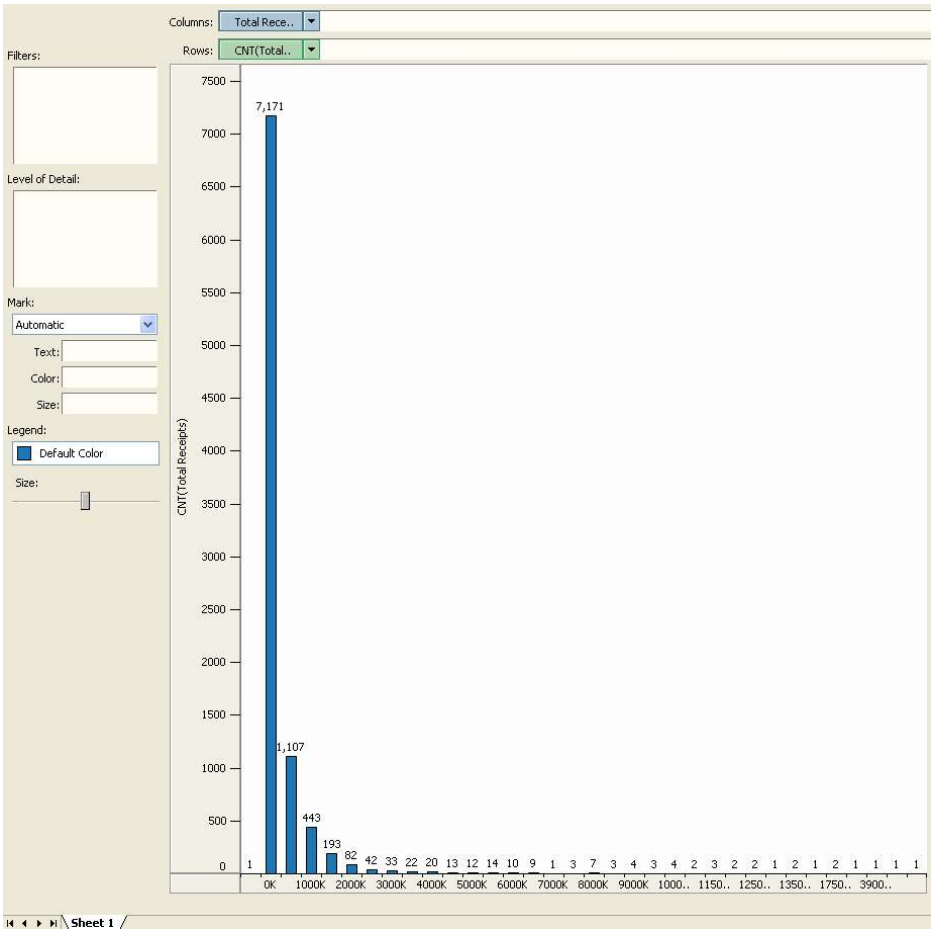

Figure 6

Histogram – number of candidates in $500,000 spending blocks

Visualizing the scatterplot alongside a bar graph showing the trend in spending over time also yielded some interesting results, thanks to Spotfire’s valuable brushing and linking features.

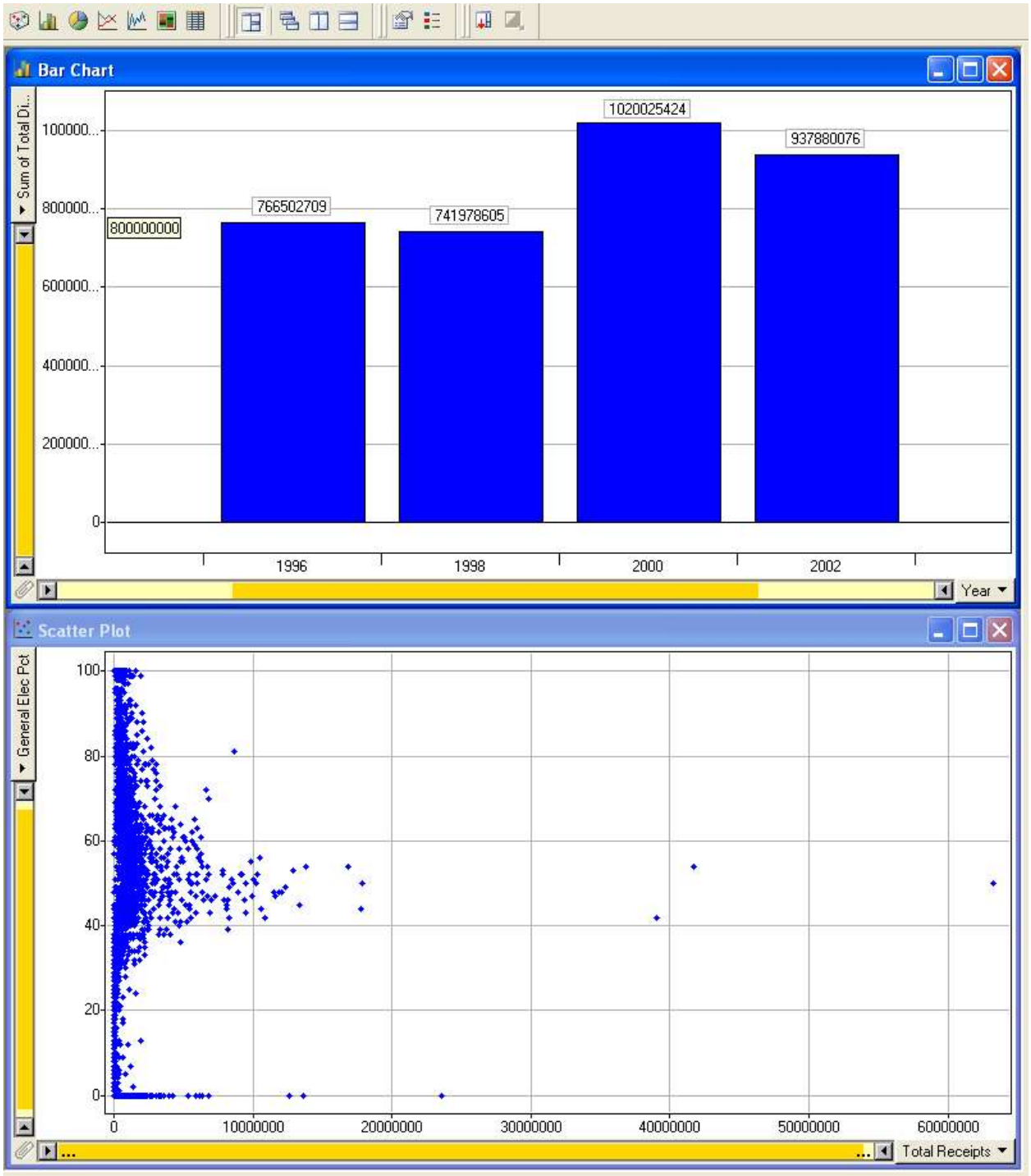

Figure 7
Total receipts by election percentage (bottom) and total receipts by year (top)

Figure 7 appears to indicate that, while overall spending 2000 was higher than any other year, the general trend is towards increased campaign spending. However, eliminating the top 10% of candidates by total receipts reveals that, for the vast majority of candidates, campaign contributions are actually decreasing. (Figure 8) Put another way, these visualizations show us

that campaign contributions are increasingly concentrated among a smaller number of candidates.

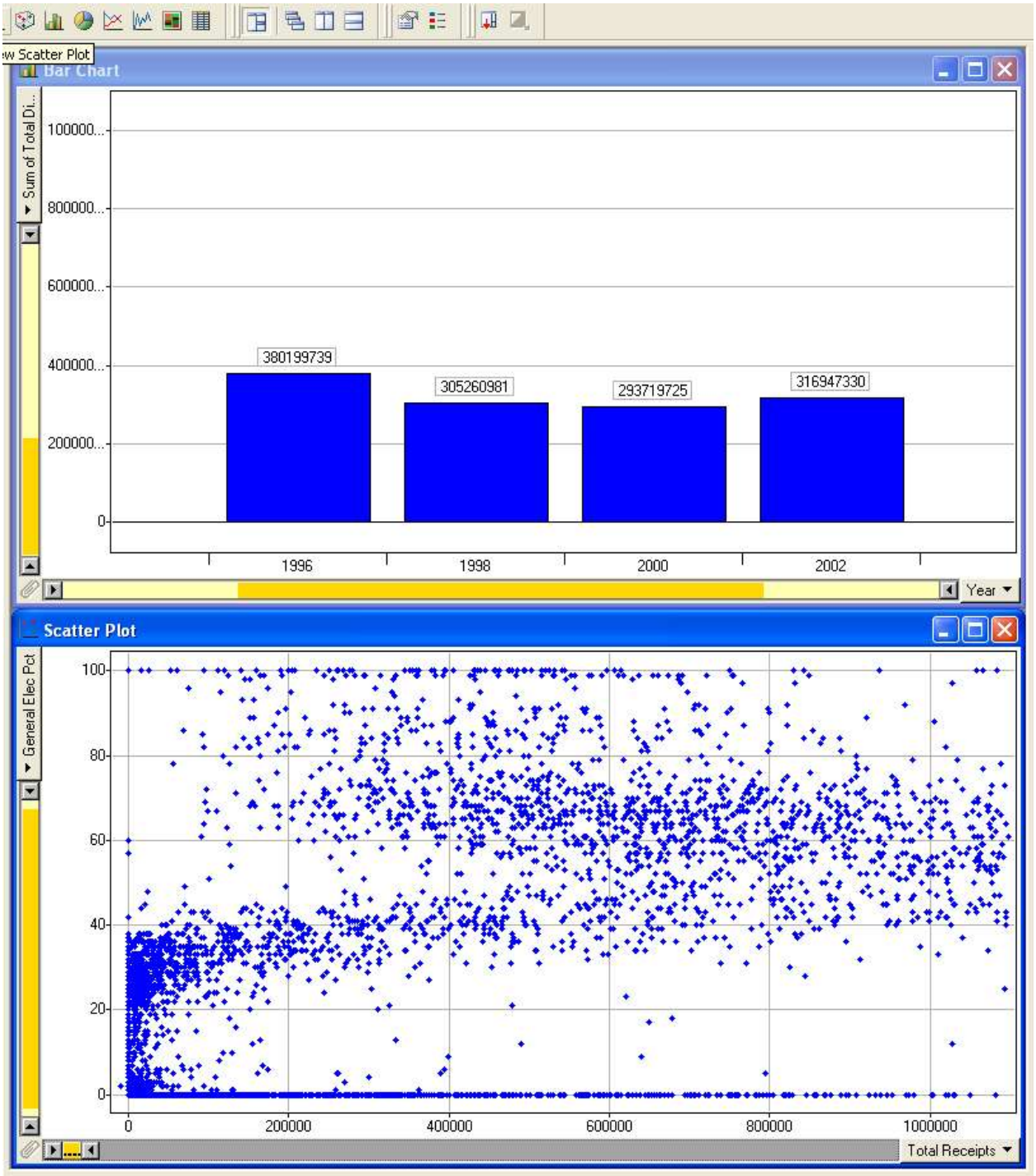

Figure 8

Total receipts by election percentage (bottom) and total receipts by year (top) with the top 10% of candidates (by total receipts) eliminated.

Revisiting the original scatterplot (Figure 5), a more subtle hypothesis emerges. There is a curious bulge along the Y axis between roughly 40 and 60 in the general election percentage. Based on the size and placement of the bulge I put forth the following hypothesis:

Campaign spending increases as general election percentage gets closer to 50% because more money is invested in races that are expected to be close.

Figure 9 shows that total receipts increase as election percentage gets closer to 50%. Figure 10 illustrates that this phenomenon holds up when we look at the most popular types of contributions across years. These visualizations seem to support the hypothesis uncovered via the initial scatterplot.

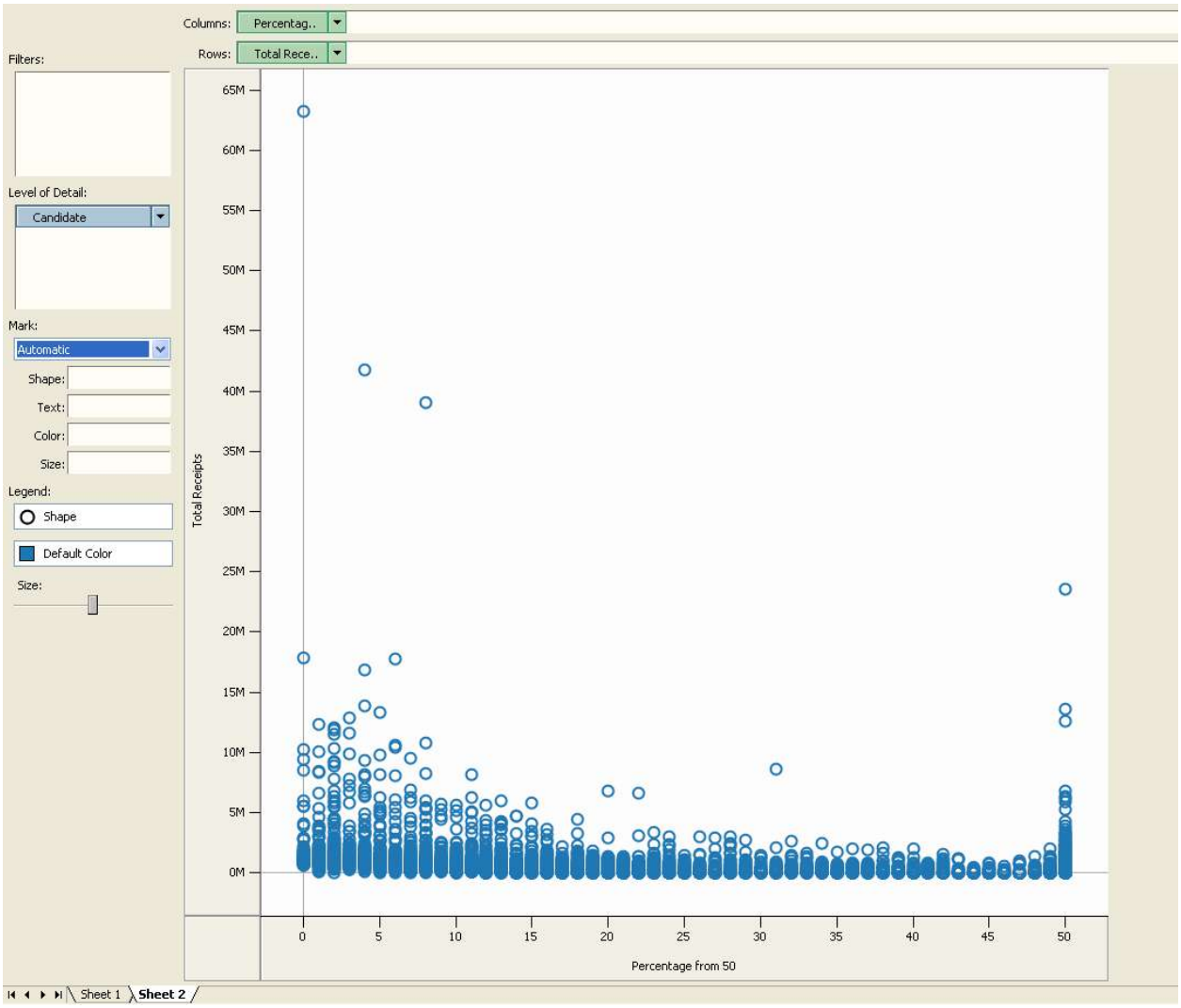

Figure 9

General election percentage proximity to 50% (x-axis) by total receipts (y-axis)

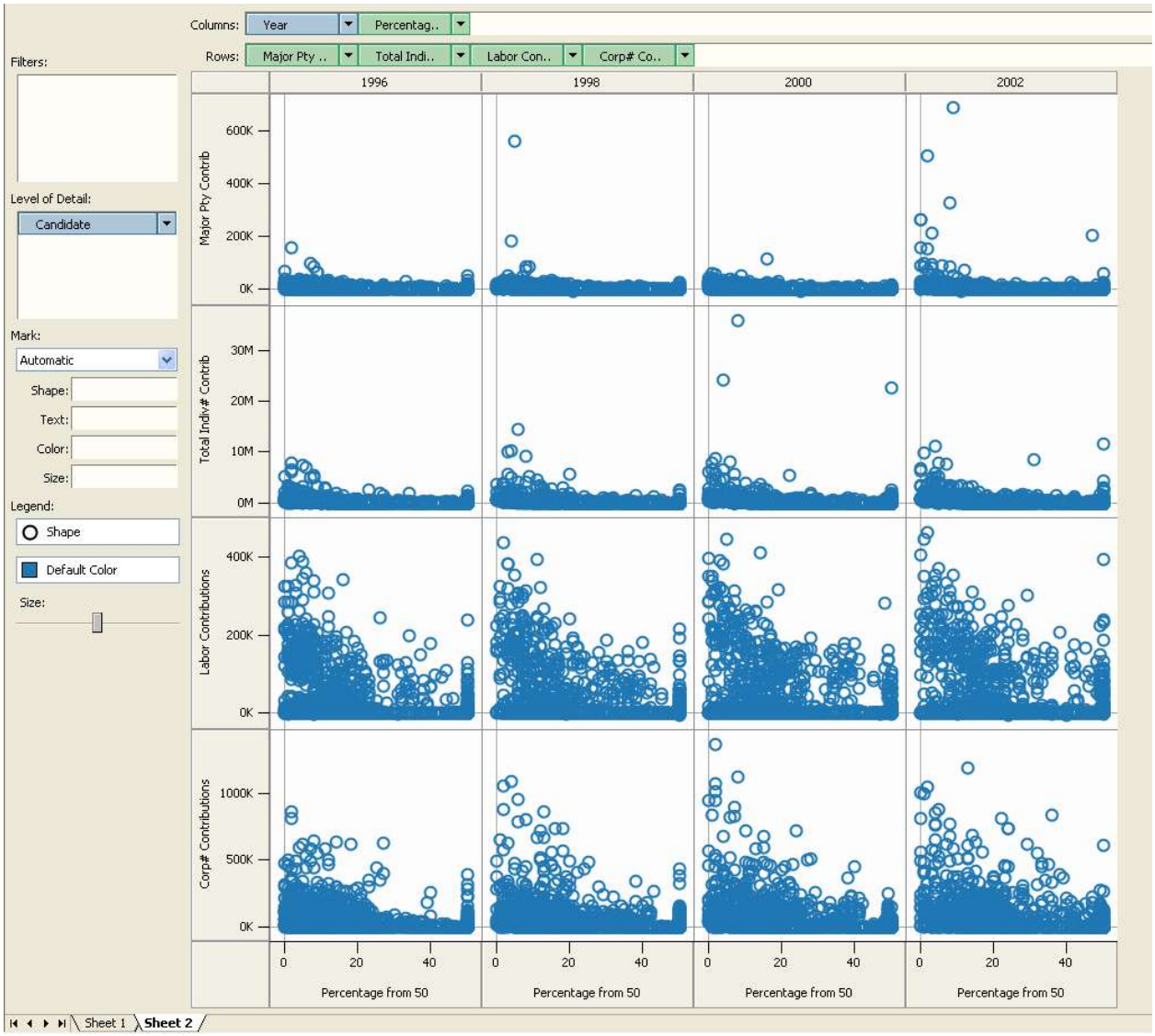

Figure 10

General election percentage proximity to 50% (x-axis) by major party contribution, total individual contribution, labor contribution, and corporate contribution (from top to bottom) across years

Each of the graphs clearly shows a kind of 'tail' – an increase in contributions for candidates whose races are not close. How can we explain this 'tail'? In order to shed more light on this end of the graph, we should filter the data according to which candidates ran in uncontested races (they received 100% of the vote) and which candidates received no votes, likely because they were ousted in the primary election.

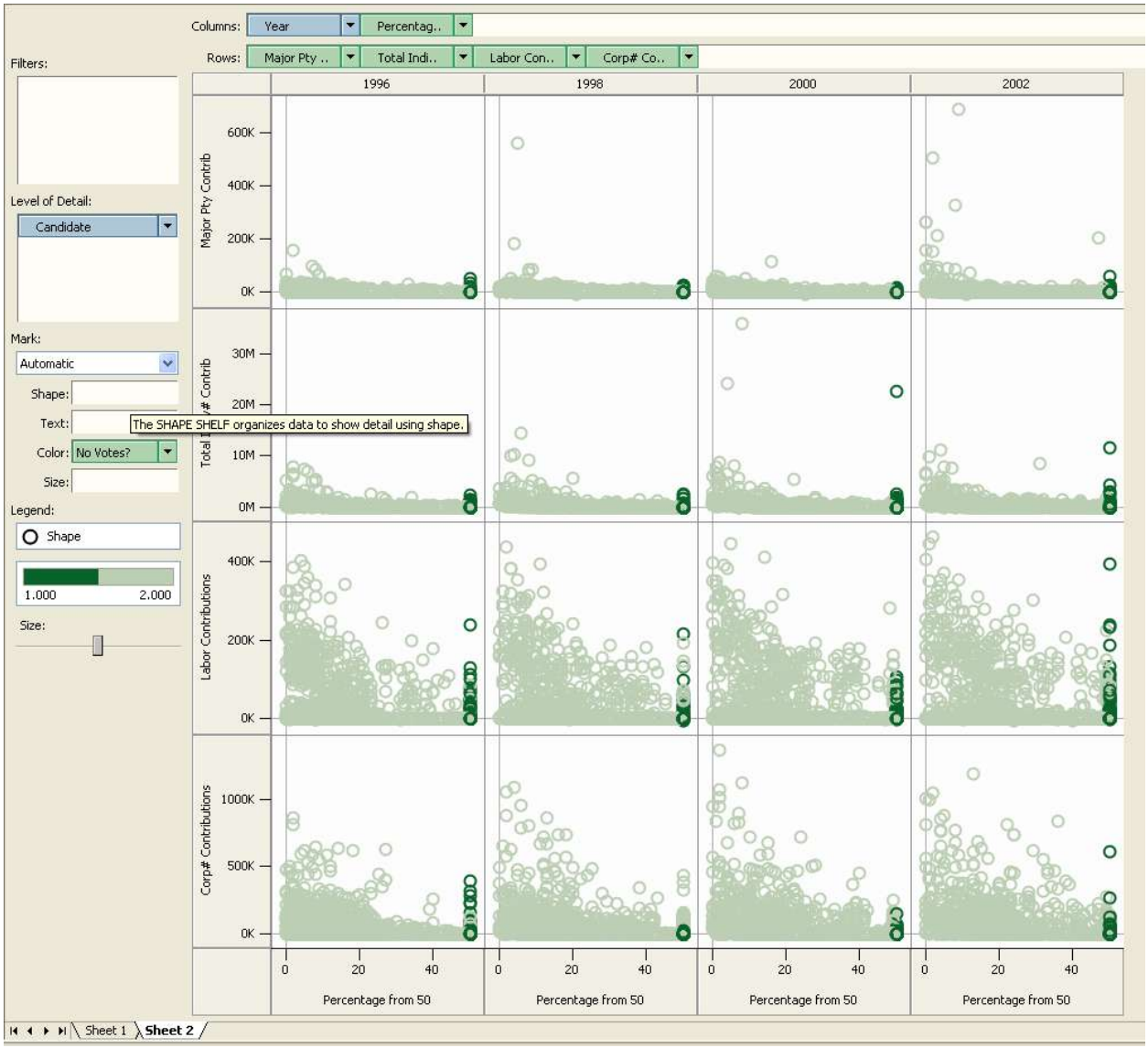

Figure 11

General election percentage proximity to 50% (x-axis) by major party contribution, total individual contribution, labor contribution, and corporate contribution (from top to bottom) across years – candidates who received 0% of votes are highlighted in dark green.

Figure 11 seems to indicate that this 'tail' is as a result of candidates who spent significantly on their campaign but lost their primary elections. If this is the case, then the 'tail' should disappear when we restrict the data to only those candidates who ran in the general election. Figure 12 confirms that the tail disappears almost entirely with respect to major party and individual contributions, but much less so with respect to labor and corporate contributions.

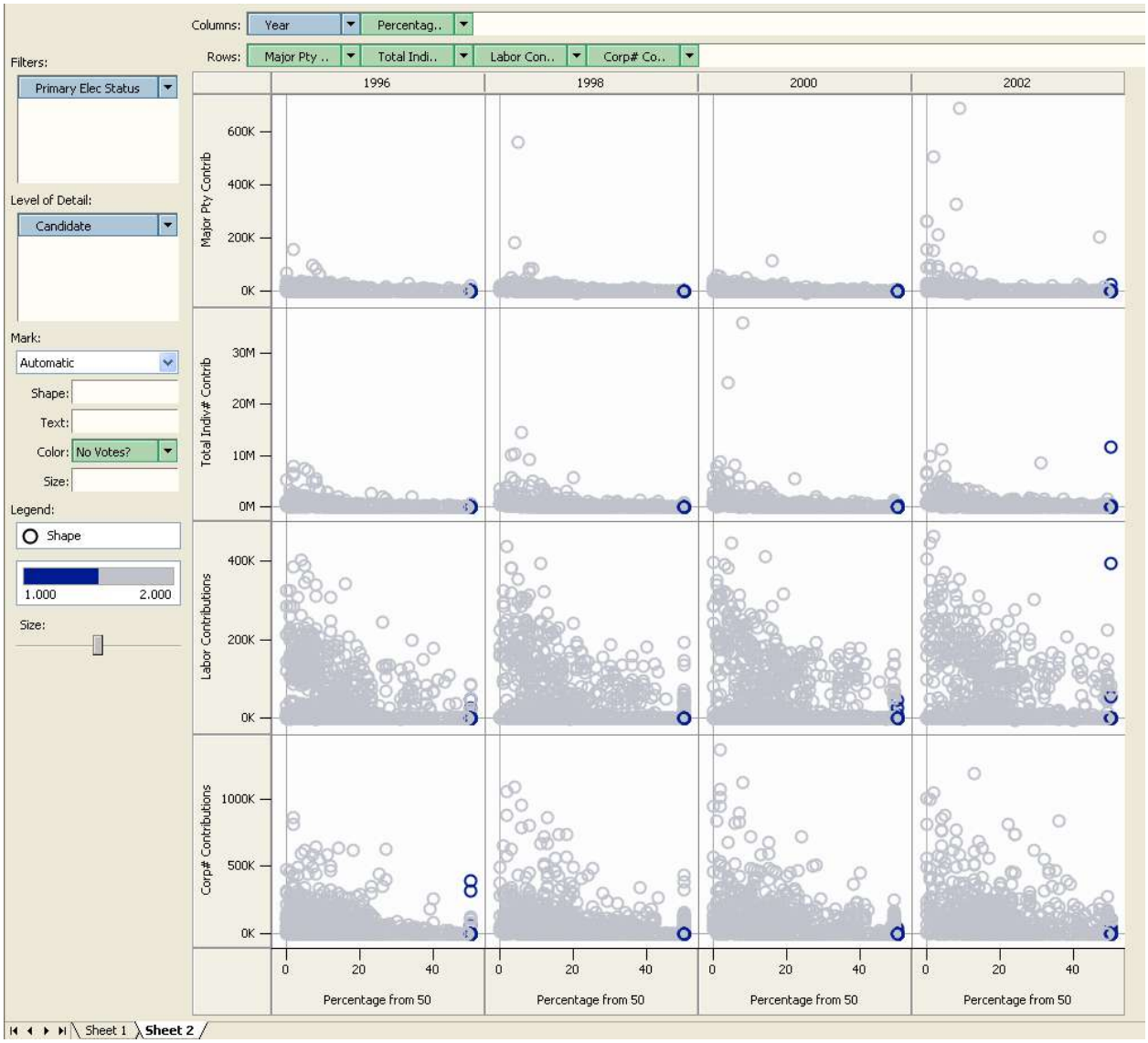

Figure 12

General election percentage proximity to 50% (x-axis) by major party contribution, total individual contribution, labor contribution, and corporate contribution (from top to bottom) across years, restricted to candidates who actually ran in the general election – candidates who received 0% of votes are highlighted in dark blue

Here we have another phenomenon which is not easily explained by the FEC data alone. While individual and major party contributions seem to track fairly well with the success of candidates – that is, contributions are higher for competitive candidates – labor and corporate contributions are much more scattered. While there remains a general trend towards funding competitive candidates, it is not concentrated in the way it is with individual and major party contributions. We might deduce from Figure 12 that labor and corporate campaign funding sources are out of touch with the realities of ongoing races, or that they are more egalitarian in their funding practices. We might also conclude that they are more idealistic, choosing to fund candidates that they support or who espouse their views regardless of their chances of winning, whereas individual and party contributors are more focused on winning elections.

Thoughts on Software

Tableau

Functionality ★★★★
Usability ★★
Aesthetics ★★★★
Overall ★★★★
(out of 5)

Pros
First and foremost, Tableau produces beautiful visualizations. Visualizations should be both telling and appealing, and Tableau provides the opportunity for both. Tableau's 'shelf' interface also makes it easy to explore the data with visualizations by swapping variables in and out of rows and columns, and quickly filtering the data and creating division by color, shape, size, etc. Of the two programs, Spotfire and Tableau, I consider Tableau to be more powerful and useful, despite the drawbacks I mention below.

Cons
Tableau is not particularly intuitive to learn. I found that the 'shelves' system for visualizing data was powerful once I understood it, but that I had to spend several hours fiddling with the software in order to understand its capabilities. In many cases, simple tasks were complicated by frustrating usability issues – in particular confusion with respect to vocabulary. I was initially unaware of the differences between and capabilities of 'dimensions' and 'measures.' In one case I wanted to divide the data by year, and was frustrated for more than 30 minutes before I stumbled upon a help document that described translating from measure to dimension. I frequently consulted the software's help system, but in general found it to be unhelpful.

Spotfire

Functionality ★★★
Usability ★★★
Aesthetics ★★
Overall ★★★
(out of 5)

Pros
While the overall package of Spotfire is less appealing than Tableau, it incorporates a number of incredibly valuable features. Spotfire's single most valuable feature is the ability to dynamically adjust the range of data using the slider bars along each axis. Especially while viewing multiple graphs with linked data, adjusting the range on one graph and watching trends change on others was fascinating. Spotfire also has useful functions for manuipulating the raw data, including the ability to create new variables based on formulas. The dedicated 'Details on Demand' window is also incredibly valuable, as it quickly allows the user to explore the data for individual records. I found this extremely useful when investigating outliers.

Cons

Spotfire’s most glaring weakness is that it does not provide easy or intuitive functions for manipulating graph annotations. Spotfire allows some discretion over where annotations are used, but does not allow the user to adjust the size or type of those annotations. Similarly, there are no options for manually editing a visualization’s legend. The default legend is small, cluttered, and full of extraneous information.