30/7/2012

Using Data Visualization to Find Insights in Data

5

This post by Gregor Aisch (Open Knowledge Foundation), is an excerpt from the Data Journalism Handbook (chapter 5: Understanding Data), freely available online under a Creative Commons Attribution-ShareAlike license.

"Visualization is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones."

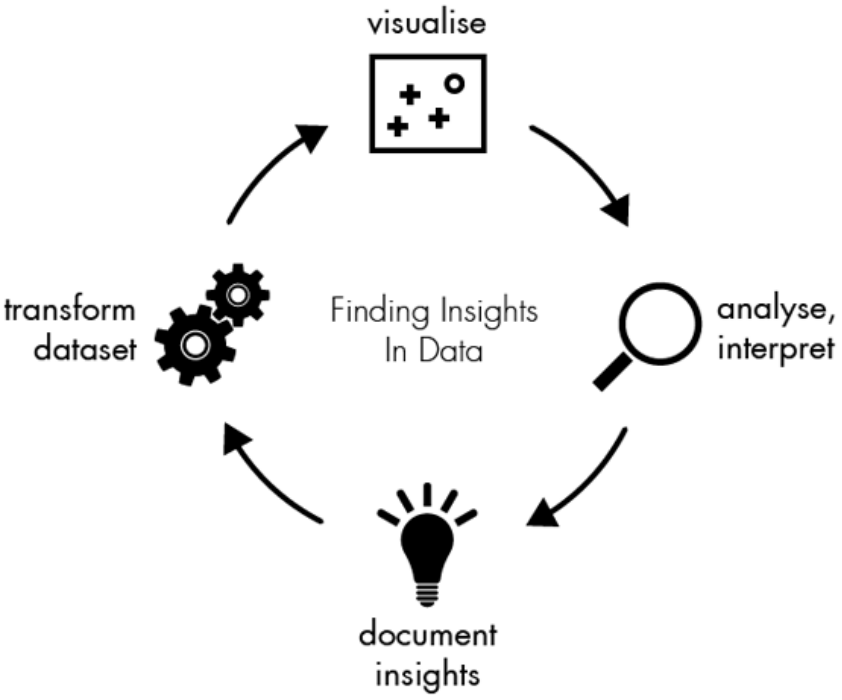
— William S. Cleveland (from Visualizing Data, Hobart Press)

Data by itself, consisting of bits and bytes stored in a file on a computer hard drive, is invisible. In order to be able to see and make any sense of data, we need to visualize it. In this section I'm going to use a broader understanding of the term "visualizing," that includes even pure textual representations of data. For instance, just loading a dataset into a spreadsheet software can be considered as data visualization. The invisible data suddenly turns into a visible "picture" on our screen. Thus, the question should not be whether journalists need to visualize data or not, but which kind of visualization may be the most useful in which situation.

In other words: when does it makes sense to go beyond the table visualization? The short answer is: *almost always*. Tables alone are definitely not sufficient to give us an overview of a dataset. And tables alone don't allow us to immediately identify patterns within the data. The most common example here are geographical patterns that can only be observed after visualizing data on a map. But there are also other kinds of patterns, which we will see later in this section.

Using Visualization to Discover Insights

It is unrealistic to expect that data visualization tools and techniques will unleash a barrage of ready-made stories from datasets. There are no rules, no "protocol" that will guarantee us a story. Instead, I think it makes more sense to look for "insights," which can be artfully woven into stories in the hands of a good journalist. Every new visualization is likely to give us some insights into our data. Some of those insights might be already known (but perhaps not yet proven), while other insights might be completely new or even surprising to us. Some new insights might mean the beginning of a story, while others could just be the result of errors in the data, which are most likely to be found by visualizing the data. In order to make finding insights in data more effective, I find the process discussed in the figure below (and the rest of this section) to be very helpful.



Data insights: a visualization (Gregor Aisch)

Learn how to visualize data

Visualization provides a unique perspective on the dataset. You can visualize data in lots of different ways. Tables are very powerful when you are dealing with a relatively small number of data points. They show labels and amounts in the most structured and organized fashion and reveal their full potential when combined with the ability to sort and filter the data. Additionally, Edward Tufte suggested including small chart pieces within table columns — for instance, one bar per row or a small line chart (since then also known as a sparkline). But still, as mentioned earlier, tables clearly have their limitations. They are great to show you one-dimensional outliers like the top 10, but they are poor when it comes to comparing multiple dimensions at the same time (for instance, population per country over time).

Major Groups	Apr-07	—	Apr-08	% YoY	% MoM	%Wt
Food	2,532		2,588	2.2	-0.3	40.9
Hospitality & Service Industries	1,159		1,195	3.1	0.1	18.7
Household Goods	858		933	8.7	0.5	13.9
Other, Pharma, Watches	552		625	13.3	-0.8	8.9
Department Stores	482		500	3.8	0.1	7.8
Clothing & Soft Goods	421		453	7.6	0.3	6.8
Recreational Goods	190		222	16.5	0.4	3.1
Total Retail Sales	6,194.2		6,515.1	5.2	-0.1	

Group	Apr-07	—	Apr-08	% YoY	% MoM	%Wt
Food						
Supermarkets & Grocery Stores	1,675		1,793	7.0	-0.3	27.0
Takeaway Food	374		271	-27.5	-2.2	6.0
Liquor	282		309	9.9	0.9	4.5
Other Food	202		215	6.4	0.8	3.3
Hospitality & Service Industries						
Hotels & Licensed Clubs	647		747	15.5	0.9	10.4
Cafes & Restaurants	456		389	-14.8	-0.7	7.4
Selected Services	56		59	5.4	-4.8	0.9
Other, Pharma, Watches						
Other Retailing	243		256	5.3	-1.9	3.9
Pharmaceutical, Cosmetic & Toiletry	216		266	23.3	0.2	3.5
Watch & Jewellery	93		103	10.6	-0.4	1.5
Department Stores						
Department Stores	482		500	3.8	0.1	7.8
Household Goods						
Furniture & Floor Covering	407		464	14.0	1.6	6.6
Domestic Hardware & Houseware	282		254	-10.0	-0.2	4.6
Domestic Appliances & Recorded Music	169		214	27.0	-1.0	2.7
Clothing & Soft Goods						
Clothing	308		331	7.6	0.4	5.0
Other Clothing Related	113		122	7.9	0.0	1.8
Recreational Goods						
Newspaper, Book & Stationery	113		139	22.7	1.2	1.8
Other Recreational Goods	77		83	7.4	-1.0	1.2

Tips from Tufte: sparklines (Gregor Aisch)

Charts, in general, allow you to map dimensions in your data to visual properties of geometric shapes. There's much written about the effectiveness of individual visual properties, and the short version is: color is difficult, position is everything. In a scatterplot, for instance, two dimensions are mapped to the to the x- and y-position. You can even display a third dimension to the color or size of the displayed symbols. Line charts are especially suited for showing temporal evolutions while bar charts are perfect for comparing categorical data. You can stack chart elements on top of each

16/6/2016 - 16/6/2016

Data Journalism Awards

The Data Journalism Awards are the first international awards recognising outstanding work in the field of data journalism worldwide.

27/6/2016 - 1/7/2016

2016 Esri User Conference (San Diego, USA)

Learn, network and share your experiences with 16,000 other Esri GIS users.

16/7/2016 - 18/7/2016

Global Summit on Computer & Information Technology (GSCIT 2016) (Tunisia)

A major platform for researchers and industry practitioners from different fields of computer and information technology.

7/8/2016 - 11/8/2016

CAR Boot Camp (Missouri, USA)

Learn how to acquire electronic information, use spreadsheets and databases to analyze the information and translate that information into high-impact stories.

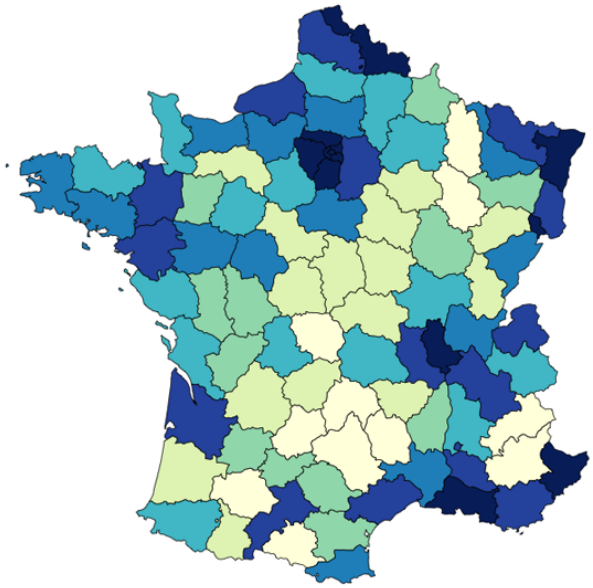
12/8/2016 - 16/8/2016

Mapping Boot Camp (Missouri, USA)

IRE and NICAR conducts this hands-on training using the latest version of ArcView GIS.

other. If you want to compare a small number of groups in your data, displaying multiple instances of the same chart is a very powerful way (also referred to as small multiples). In all charts you can use different kinds of scales to explore different aspects in your data (e.g., linear or log scale).

In fact, most of the data we're dealing with is somehow related to actual people. The power of maps is to re-connect the data to our very physical world. Imagine a dataset of geo-located crime incidents. Crucially, you want to see *where* the crimes happen. Also maps can reveal geographic relations within the data, e.g. a trend from North to South or from urban to rural areas.



Choropleth Map (Gregor Aisch)

Speaking of relations, the fourth most important type of visualization is a graph. Graphs are all about showing the interconnections (edges) in your data points (nodes). The position of the nodes is then calculated by more or less complex graph layout algorithms which allow us to immediately see the structure within the network. The trick of graph visualization in general is to find a proper way to model the network itself. Not all datasets already include relations, and even if they do, it might not be the most interesting aspect to look at. Sometimes it's up to the journalist to define edges between nodes. A perfect example of this is the U.S. Senate Social Graph, whose edges connect senators that voted the same in more than 65% of the votes.

Analyze and interpret what you see
Once you have visualized your data, the next step is to learn something from the picture you created. You could ask yourself:

- What can I see in this image? Is it what I expected?
- Are there any interesting patterns?
- What does this mean in the context of the data?

Sometimes you might end up with a visualization that, in spite of its beauty, might seem to tell you nothing of interest about your data. But there is almost always something that you can learn from any visualization, however trivial.

Document your insights and steps
If you think of this process as a journey through the dataset, the documentation is your travel diary. It will tell you where you have traveled to, what you have seen there, and how you made your decisions for your next steps. You can even start your documentation before taking your first look at the data. In most cases when we start to work with a previously unseen dataset, we are already full of expectations and assumptions about the data.

Usually there is a reason why we are interested in that dataset that we are looking at. It's a good idea to start the documentation by writing down these initial thoughts. This helps us to identify our bias and reduces the risk of misinterpretation of the data by just finding what we originally wanted to find.

I really think that the documentation is the most important step of the process—and it is also the one we're most likely to tend to skip. As you will see in the example below, the described process involves a lot of plotting and data wrangling. Looking at a set of 15 charts you created might be very confusing, especially after some time has passed. In fact, those charts are only valuable (to you or any other person you want to communicate your findings) if presented in the context in which they

have been created. Hence you should take the time to make some notes on things like:

- Why have I created this chart?
- What have I done to the data to create it?
- What does this chart tell me?

Transform data
Naturally, with the insights that you have gathered from the last visualization, you might have an idea of what you want to see next. You might have found some interesting pattern in the dataset which you now want to inspect in more detail. Possible transformations are:

Zooming
To have look at a certain detail in the visualization

Aggregation
To combine many data points into a single group

Filtering
To (temporarily) remove data points that are not in our major focus

Outlier removal
To get rid of single points that are not representative for 99% of the dataset.

Let's consider that you have visualized a graph, and what came out of this was nothing but a mess of nodes connected through hundreds of edges (a very common result when visualizing so-called *densely connected networks*). One common transformation step would be to filter some of the edges. If, for instance, the edges represent money flows from donor countries to recipient countries, we could remove all flows below a certain amount.

Which Tools to Use

The question of tools is not an easy one. Every data visualization tool available is good at something. Visualization and data wrangling should be easy and cheap. If changing parameters of the visualizations takes you hours, you won't experiment that much. That doesn't necessarily mean that you don't need to learn how to use the tool. But once you learned it, it should be really efficient. It often makes a lot of sense to choose a tool that covers both the data wrangling and the data visualization issues. Separating the tasks in different tools means that you have to import and export your data very often. Here's a short list of some data visualization and wrangling tools:

- Spreadsheets like LibreOffice, Excel or Google Docs
- Statistical programming frameworks like R (r-project.org) or Pandas (pandas.pydata.org)
- Geographic Information Systems (GIS) like Quantum GIS, ArcGIS, or GRASS
- Visualization Libraries like d3.js (mbostock.github.com/d3), Prefuse (prefuse.
- org), or Flare (flare.prefuse.org)
- Data wrangling tools like Google Refine or Datawrangler
- Non-programming visualization software like ManyEyes or Tableau Public (tableausoftware.
- com/products/public)

The sample visualizations in the next section were created using R, which is kind of a Swiss Army knife of (scientific) data visualization.

An Example: Making Sense of US Election Contribution Data

Let us have look at the US Presidential Campaign Finance database, which contains about 450,000 contributions to US presidential candidates. The CSV file is 60 megabytes and way too big to handle easily in a program like Excel.

In the first step I will explicitly write down my initial assumptions on the FEC contributions dataset:

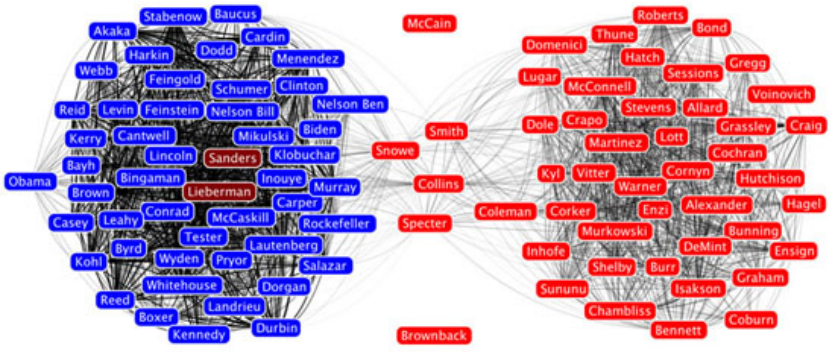
- Obama gets the most contributions (since he is the president and has the greatest
- popularity)
- The number of donations increases as the time moves closer to election date
- Obama gets more small donations than Republican candidates

To answer the first question, we need to *transform* the data. Instead of each single contribution, we need to sum the total amounts contributed to each candidate. After *visualizing* the results in a sorted table, we can confirm our assumption that Obama would raise the most money:

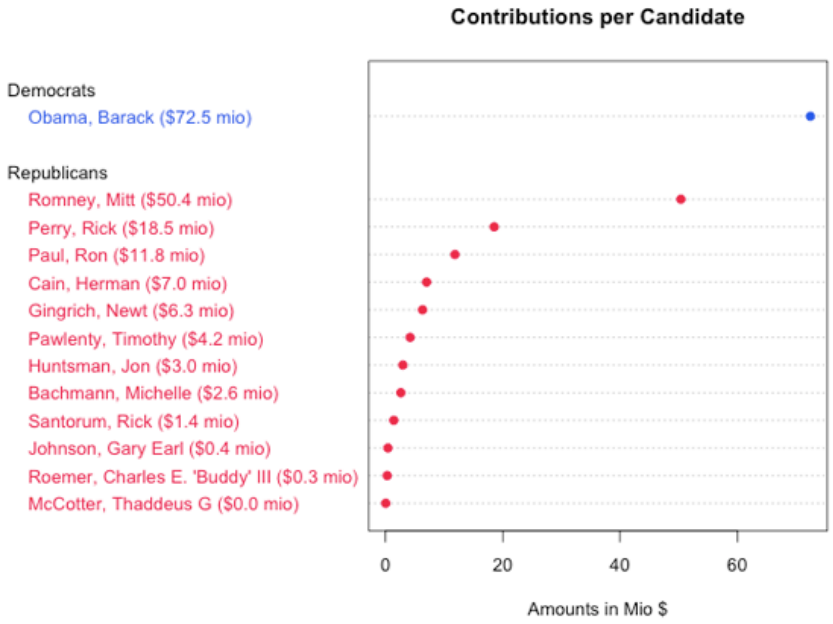
Candidate	Amount ($)
Obama, Barack	72,453,620.39
Romney, Mitt	50,372,334.87
Perry, Rick	18,529,490.47
Paul, Ron	11,844,361.96
Cain, Herman	7,010,445.99
Gingrich, Newt	6,311,193.03
Pawlenty, Timothy	4,202,769.03
Huntsman, Jon	2,955,726.98
Bachmann, Michelle	2,607,916.06
Santorum, Rick	1,413,552.45
Johnson, Gary Earl	413,276.89
Roemer, Charles E. *Buddy* III	291,218.80
McCotter, Thaddeus G	37,030.00

Even though this table shows the minimum and maximum amounts and the order, it does not tell very much about the underlying patterns in candidate ranking. Figure 75 is another view on the data, a chart type that is called 'dot chart' in which we can see everything that is shown in the table plus the patterns within the field. For instance, the dot chart allows us to immediately compare the distance between Obama and Romney and Romney and Perry without needing to subtract values. (Note: The dot chart was created using R. You can find links to the source codes at the end of this chapter).

Now, let us proceed with a bigger picture of the dataset. As a first step, I visualized all contributed amounts over time in a simple plot. We can see that almost all donations are very, very small compared to three really big outliers. Further investigation reveals that these huge contributions are coming from the "Obama Victory Fund 2012" (also known as Super PAC) and were made on June 29th ($450k), September 29th ($1.5mio), and December 30th ($1.9mio).

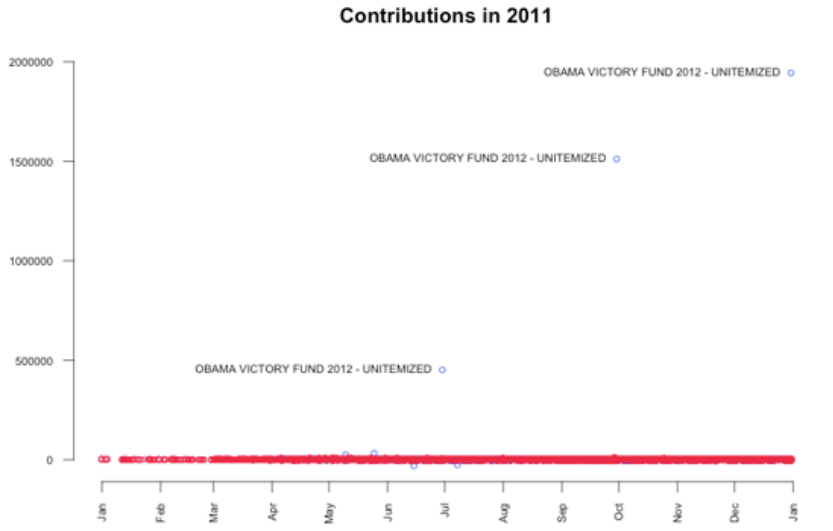


Visualizations to spot underlying patterns (Gregor Aisch)



Three clear outliers (Gregor Aisch)

While the contributions by Super PACs alone is undoubtedly the biggest story in the data, it might be also interesting to look beyond it. The point now is that these big contributions disturb our view on the smaller contributions coming from individuals, so we're going to remove them from the data. This transform is commonly known as outlier removal. After visualizing again, we can see that most of the donations are within the range of $10k and -$5k.



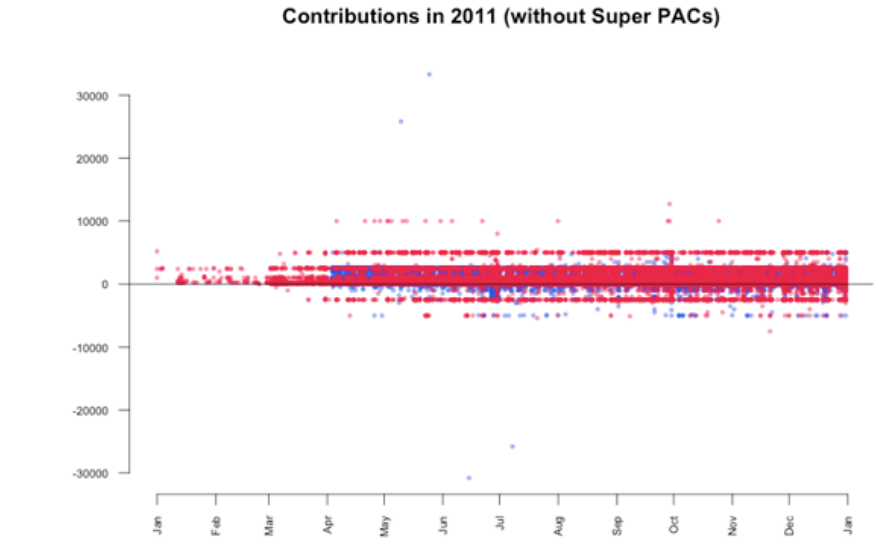
Removing the outliers (Gregor Aisch)

According to the contribution limits placed by the FECA, individuals are not allowed to donate more than $2500 to each candidate. As we see in the plot, there are numerous donations made above that limit. In particular, two big contributions in May attract our attention. It seems that they are *mirrored* in negative amounts (refunds) in June and July. Further investigation in the data reveals the following transactions:

- On May 10, *Stephen James Davis*, San Francisco, employed at Banneker Partners (attorney), has donated **$25,800** to Obama.
- On May 25, *Cynthia Murphy*, Little Rock, employed at the Murphy Group (public relations), has donated **$33,300** to Obama.
- On June 15, the amount of **$30,800** was refunded to *Cynthia Murphy*, which reduced the donated amount to **$2500**.
- On July 8, the amount $25,800 was refunded to Stephen James Davis, which reduced the donated amount to $0.

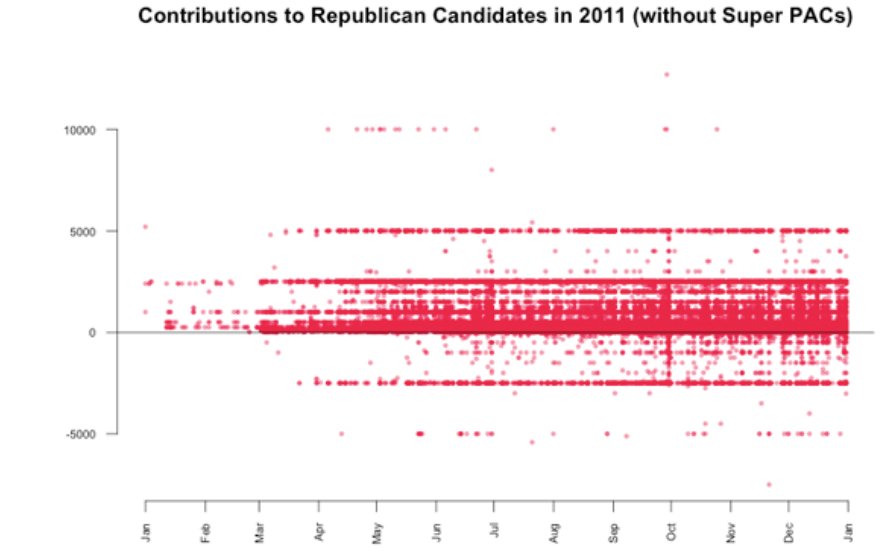
What's interesting about these numbers? The $30,800 refunded to Cynthia Murphy equals the

maximum amount individuals may give to national party committees per year. Maybe she just wanted to combine both donations in one transaction, which was rejected. The $25,800 refunded to Stephen James Davis possibly equals the $30,800 minus $5000 (the contribution limit to any other political committee). Another interesting finding in the last plot is a horizontal line pattern for contributions to Republican candidates at $5000 and -$2500. To see them in more detail, I visualized just the Republican donations. The resulting graphic is one great example of patterns in data that would be invisible without data visualization.



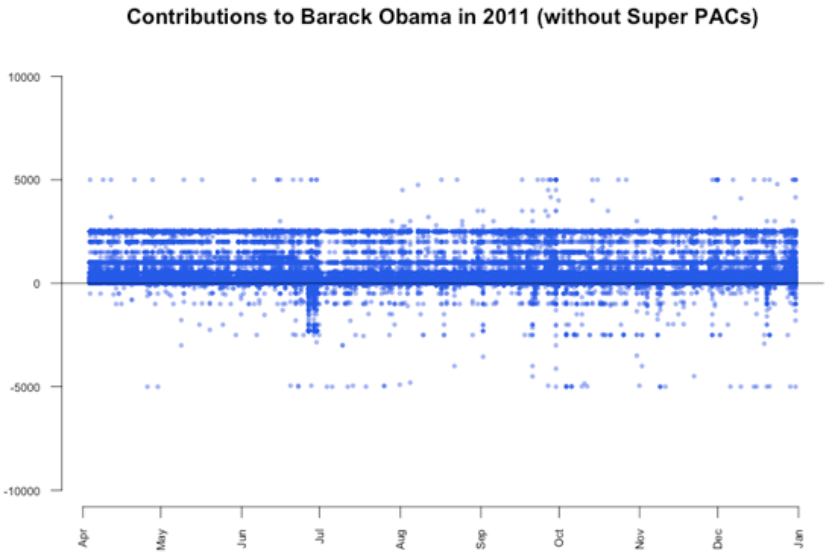
Removing outliers 2 (Gregor Aisch)

What we can see is that there are many $5000 donations to Republican candidates. In fact, a look up in the data returns that these are 1243 donations, which is only 0.3% of the total number of donations, but since those donations are evenly spread across time, the line appears. The interesting thing about the line is that donations by individuals were limited to $2500. Consequently, every dollar above that limit was refunded to the donors, which results in the second line pattern at -$2500. In contrast, the contributions to Barack Obama don't show a similar pattern.



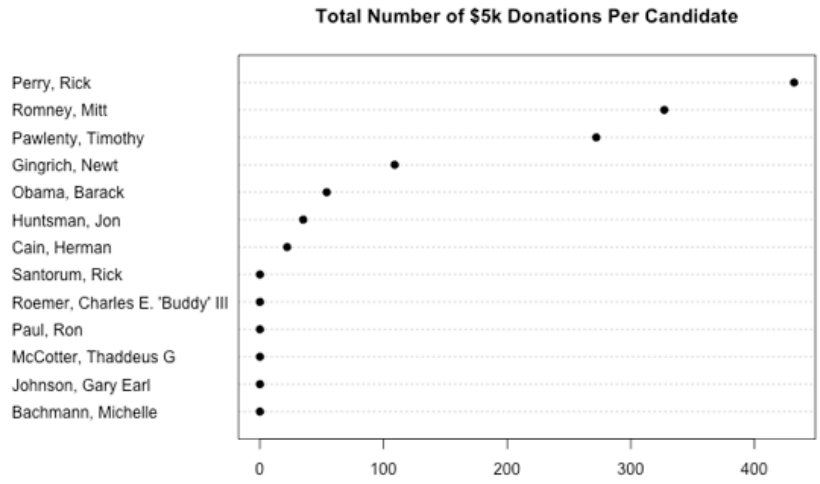
Removing outliers 3 (Gregor Aisch)

So, it might be interesting to find out why thousands of Republican donors did not notice the donation limit for individuals. To further analyze this topic, we can have a look at the total number of $5k donations per candidate.



Donations per candidate (Gregor Aisch)

Of course, this is a rather distorted view since it does not consider the total amounts of donations received by each candidate. The next plot shows the percentage of $5k donations per candidate.



Where does the senator's money come from?: donations per candidate (Gregor Aisch)

What To Learn From This

Often, such a visual analysis of a new dataset feels like an exciting journey to an unknown country. You start as a foreigner with just the data and your assumptions, but with every step you make, with every chart you render, you get new insights about the topic. Based on those insights, you make decisions for your next steps and what issues are worth further investigation. As you might have seen in this chapter, this process of visualizing, analyzing and transformation of data could be repeated nearly infinitely.

Get the Source Code

All of the charts shown in this chapter were created using the wonderful and powerful software R. Created mainly as a scientific visualization tool, it is hard to find any visualization or data wrangling technique that is not already built into R. For those who are interested in how to visualize and wrangle data using R, here's the source code of the charts generated in this chapter:

- dotchart: contributions per candidate
- plot: all contributions over time
- plot: contributions by authorized committees

There is also a wide range of books and tutorials available.

Return to Resources overview

Comments

Andy · 4 years ago

Well done Guest (is that your real name?), excellent use of your time and keyboard and an invaluable contribution to the article. To answer your first question, you would also use data visualisation to communicate/convey data stories and insights to others in an explanatory sense. By contrast, Gregor has skilfully described how you use it in an Exploratory context. These are two very different situations for using data visualisation. To answer your second question (ie. "Dur?") I'm going to need more words...

Guest · 4 years ago

What else would you use data visualisation for besides finding insights in data? Dur?

Ministry of Education, Culture and Science

Data driven journalism was created by the European Journalism Centre (EJC) with partial funding from the Dutch Ministry of Education, Culture and Science.