Recreating News Visualizations in Tableau

Robert Kosara, Tableau Research

October 2015

Minor updates: Tamara Munzner, 12 October 2015

This guide walks you through the steps to create Tableau visualizations that mimic some particularly good news graphics. The goal is to show what can be done in Tableau, and to show/introduce some of its features in a concrete use context.

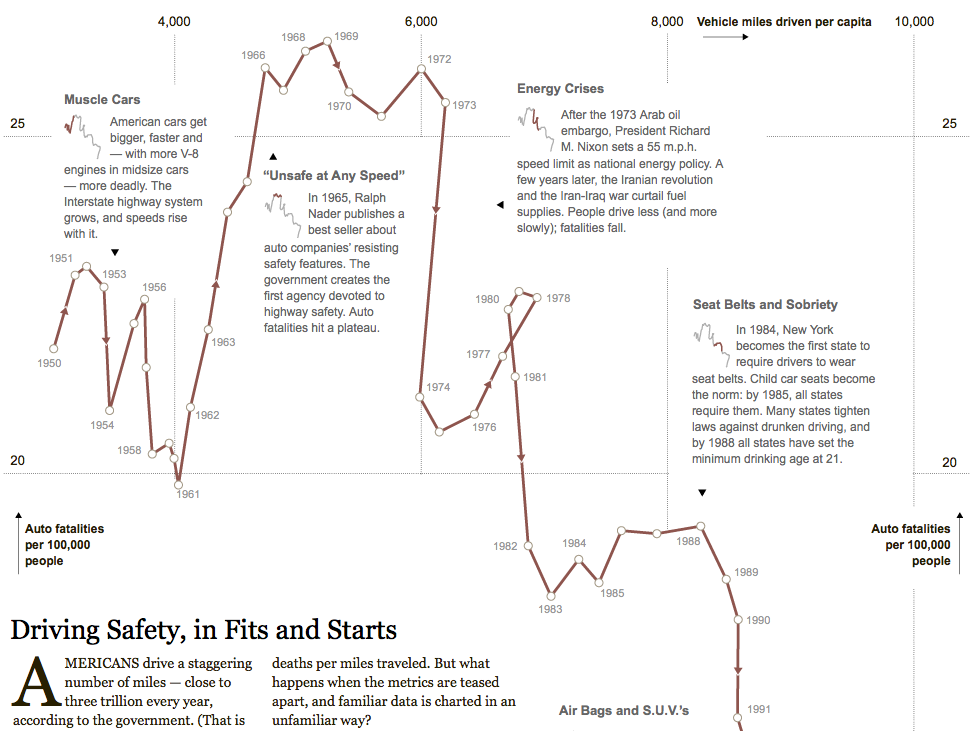
Connected Scatterplot with Annotations 1

Small Multiples Area Chart of Drought Index with Decade Breaks 4

Case-Shiller Home Price Index with Highlighting and Custom Index Date 7

# Connected Scatterplot with Annotations

The connected scatterplot is an interesting technique for comparing data over time. The inspiration for this comes from one of my favorite news graphics ever, though we’ll use different data. The piece is *Driving Safety, in Fits and Starts* by Hannah Fairfield for the New York Times (September 17, 2012): <http://nyti.ms/PB07e2>



The connected scatterplot is easy to create but can be very interesting to explore. Instead of the data used above, though, we’ll look at the minimum wage in the U.S.

Create a new sheet and pick the Minimum Wage 2015 data source. Then follow these steps to create a first chart of the nominal minimum wage (the amount of dollars people are paid at any particular time) and the actual buying power in constant dollars (so we can compare what those nominal values actually mean).

* Drag *Date* to the *Columns* shelf
* Drag *Nominal* to *Rows*
* Click on *Year(Date)* and change the detail level to *Month* in the second list, the pill will turn green
* Click on the green “pill” in the *Rows* shelf that says *SUM(Nominal)* and from the *Measure (Sum)* menu, pick *Maximum* instead
  + This is because this particular dataset has two entries for months when the minimum wage changed, which is causing little bumps that aren’t real.
* Now drag *Real* to the *Rows* shelf as well
* Make the same change from *Sum* to *Maximum* for the aggregation here
* Tableau has created two separate line charts, but we want both lines in the same chart. So click on the *Max(Real)* pill and pick *Dual Axis* from the menu.
* Now we have both lines, but they’re on different axes (compare the values on the far left to the far right). Right-click the axis on the right and select *Synchronize Axis.*

This is not the connected scatterplot, but it will be a useful additional view to use in a moment.

To create the connected scatterplot…

* Right-click on the tab and select *Duplicate Sheet* from the menu
* Drag *MONTH(Date)* from *Columns* to the *Detail* field on the Marks card
* Drag *MAX(Nominal)* to *Columns*
* This gives us a scatterplot, but we want to connect the dots. Click on the dropdown that says *Automatic* on the Marks card and change it to *Line*.
* Drag the *MONTH(Date)* field from the bottom of the Marks card onto the *Path* field while holding down the command (**⌘**) key on the Mac or Ctrl on Windows. This adds the field to the path in addition to where it already is, rather than replacing its function. The path makes sure the points are connected in the right order.
* You can drag off the *Measure Names* from the Marks card. This was generated by Tableau for the dual-axis chart above, but it’s not needed here.
* Click on the color field and pick black or dark gray to make the view look more like print

To put the views together, we’ll create a dashboard.

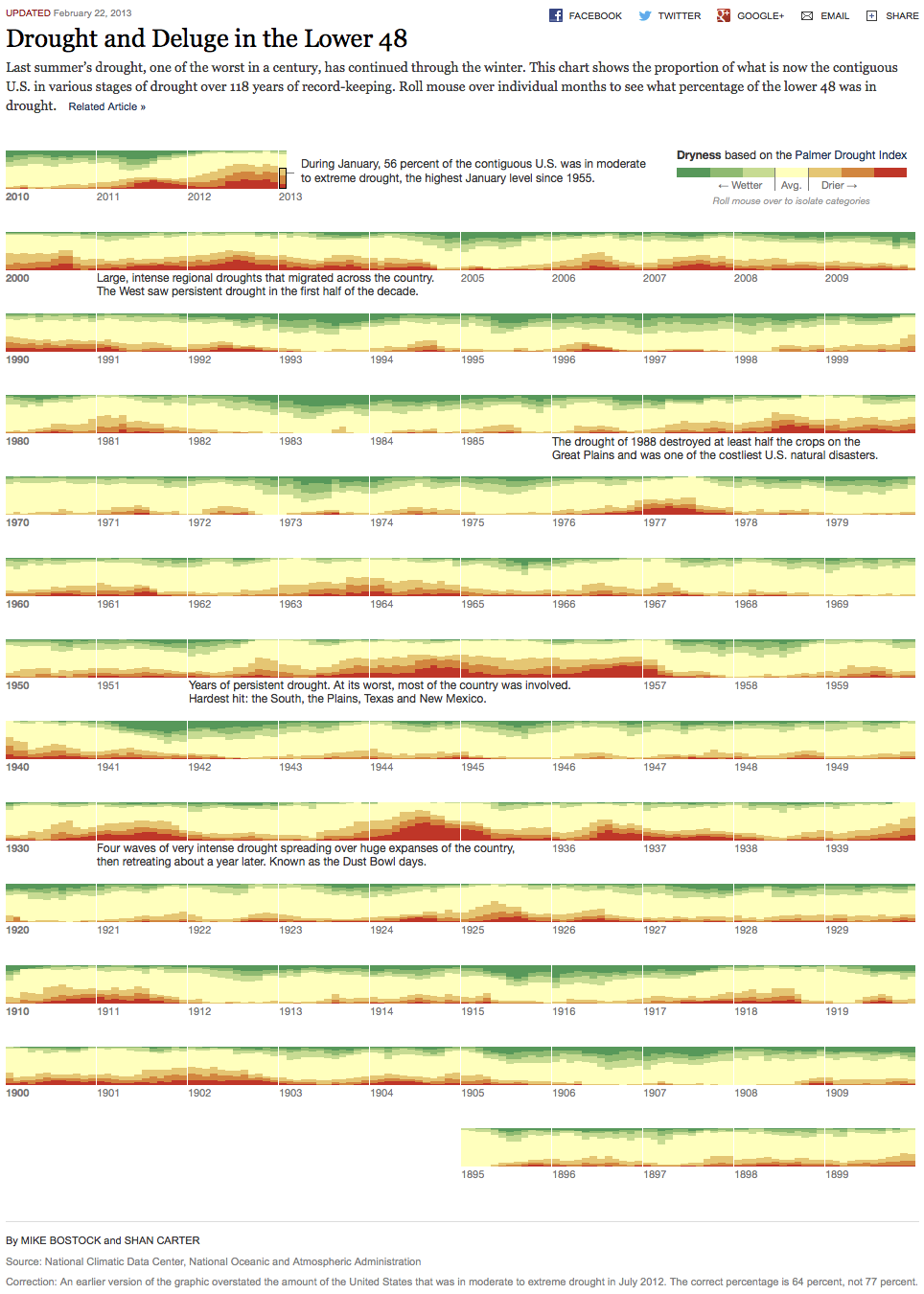
* Right-click on the tab and select *New Dashboard* from the menu, or use the middle button in the lower right of the window
* A dashboard lets you combine multiple worksheets, set up interactions between them, etc. For this exercise, we’ll just use it for layout.
* There is a list of worksheets in the upper left. Drag the last one you just created into the main window and drop it there (there’s a preview that pops up as you mouse over to help you pick).
* Now start dragging the sheet with the dual-axis chart, but hold down the shift key before dropping it. That makes it a so-called floating zone that does not try to modify the underlying layout.
* You can move floating items around by grabbing their top bars and resize them in a way similar to windows
* Drag out a *Text* object from the left to create text boxes that you can edit
* Turn off the titles of the two charts by clicking on the triangles in the upper right and unchecking that option in the menus
* Change axis labels (and hide axes) by right-clicking and double-clicking

The steps to a newspaper-style layout are hard to describe but are not very difficult to figure out from the image below.



# Small Multiples Area Chart of Drought Index with Decade Breaks

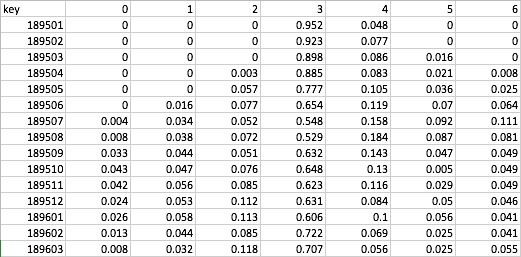
*Drought and Deluge in the Lower 48* is a great example of a small multiples plot: <http://nyti.ms/Obq0SH>



The data is already pre-loaded in the Drought Data data source in the workbook.

For practice in data wrangling, try recreating it from the raw data that you can scrape from the web site above: reading the page's HTML source shows that the data lives at  [http://graphics8.nytimes.com/newsgraphics/2013/02/22/drought-timeline/f94f53e687f19dd1cd10a6adb8fff03846d2a532/pdsi.tsv](http://graphics8.nytimes.com/newsgraphics/2013/02/22/drought-timeline/f94f53e687f19dd1cd10a6adb8fff03846d2a532/pdsi.tsv" \t "_blank)

This is what that looks like after being pasted into Excel:



The key field is the date, with the first four digits being the year and the last two the month. We’ll deal with that in a moment. The columns are for one of the drought severeness levels each, from 0 to 6. Simply pivot those columns and then rename the first column to *Level* and the second one to *Amount*.

* Once on the sheet, right-click in the data pane and select *Create Calculated Field…*
* Name it Date and enter this expression: DATEPARSE("yyyyMM", [Key])
  + The dateparse() function uses a pattern to match a date it can’t otherwise understand. The first part, *yyyy*, is a four-digit year; *MM* stands for a numerical month (as opposed to the month name).
* Click OK to close the formula editor

After wrangling, to build the initial view, follow these steps:

* Drag *Date* to the *Columns* shelf
* Drag *Amount* to *Rows*
* Drag *Level* to *Color*
* In the dropdown at the top of the *Marks* card, select *Bar*
* Click on the little + button in the *Date* field in the *Columns* shelf, then click that also on the *Quarter* pill that appears
* Drag the quarter off to an empty area to remove it from the view
* In the *Amount* pill on the Rows shelf, click the little arrow and from *Quick Table Calculation,* select *Percent of Total*
* Then click that again and from *Compute Using*, select *Cell*
* This makes sure all the bar stacks are the same height. Since the numbers are cut off after three decimals, they might not always sum to 1 exactly.

Now we have the initial view, but it’s upside down, with the more severe drought levels (higher numbers) at the bottom. We can fix this by dragging the items in the color legend so that the highest number is on top and the lowest on the bottom.

Also, click on the Size field and adjust the width of the bars until there is no more visible white space between bars within a year. This creates the nice, even appearance the NY Times piece has.

All the years are now laid out next to each other. If you scroll to the right, you will see them all. But we want a more tabular layout (called a *trellis*), with ten years in each row. To do this, we need to help Tableau a little. We’ll calculate two numbers for each year: the decade it is in and how far it is away from the start of the decade.

* Right-click in the data pane and create a new calculated field
* Name it *Decade* and enter the following expression: FLOOR(DATEPART('year',[Date])/10)\*10
  + The DATEPART() function extracts a particular element from a date, in this case the year
  + We divide that number by 10 and round the result of that down using the FLOOR() function
  + For 1956, the result of that is FLOOR(1956/10)=FLOOR(195.6)=195
  + To make the decade number nicer, we then multiply by ten again, which for the 1956 example gives us 195\*10=1950
* That is the first part of our grid (the rows), but we also need to order the years within each decade. Create a calculated field called *Decade Offset*
* Enter the expression DATEPART('year', [Date])-[Decade]
  + This takes the year and subtracts the decade. For the 1956 example, that would be 1956-1950=6
* Before we can use the fields, we will need to tell Tableau that we don’t want to treat these as continuous numbers. To do that, drag *Decade* and *Decade Offset* from the *Measures* area in the data pane to the *Dimensions*.
* Now drop the *Decade* on the *Rows* shelf
* Drop the *Decade Offset* on the *Columns* shelf to replace the year (either drop on the year or remove the year once the offset is there)
* This shows the 1890s at the top. To change that, mouse over the Decade axis label and click the little sort button that appears.

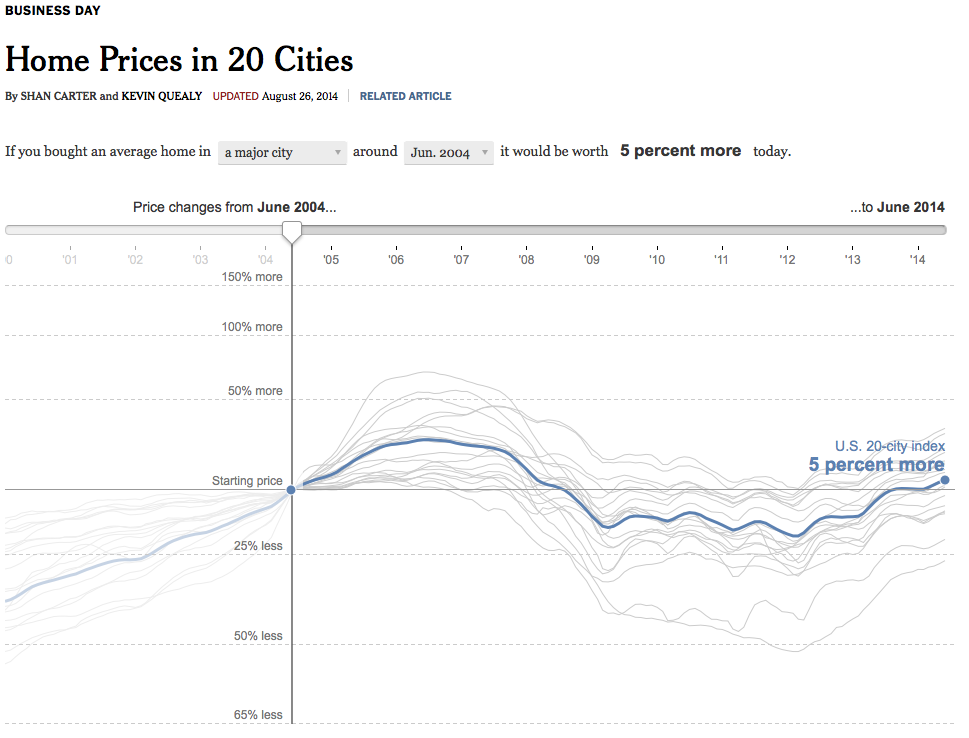
That is the basic layout. We can now remove the row headers by right-clicking on the little axis label saying *% of Total Amount* and deselecting *Show Header*.

A variation of this, based on different data, would be maps over time like in this example (available online at <http://nyti.ms/LwehKV)>:



# Case-Shiller Home Price Index with Highlighting and Custom Index Date

We will recreate this piece from the NY Times: <http://nyti.ms/NUUD2e>



The data is coming directly from the source. The Case-Shiller Index measures the changes in home price in 20 major U.S. cities, and there are composite indices for 10 and 20 cities, as well the U.S. overall.

The data lives here: <http://us.spindices.com/index-family/real-estate/sp-case-shiller>

To download it, click on the *Additional Info* dropdown and select *Home Price Index Levels* or *Seasonally Adjusted Home Price Index Levels*.

First, reshape the data so that it looks like this. I won’t cover that part in detail here, but it basically requires the following steps:

* Rename *Composite-10*, *Composite-20*, and *National-US* to *US-Composite 10*, *US-Composite 20*, and *US-National*, respectively.
* Select all columns but the first and perform a pivot
* Split the field “Pivot field names”
* Rename the first column to *Date*, the third one to *State*, the fourth to *City*, the fifth to *Index*
* Change the type of the last column to Number (Decimal)
* Select the appropriate geographic roles for City and State

Now create the initial view

* Drag *Date* onto the *Columns* shelf
* Drag *Index* onto *Rows*
* Drag *City* onto *Color* (in the Marks card)
* Click on *Date* in the Columns shelf to switch it to continuous *Month* (the *Month* item in the second list in the menu)

That is the initial view. But we don’t want all of this. The index is based on the prizes in each city on January 1, 2000. So let’s only look at data from that date forward.

* Drag *Date* onto the *Filters* card
* In the dialog that pops up, select *Range of Dates*
* On the next page, click on the *Starting Date* field at the top
* Click on the little calendar symbol in the date field and navigate to January 2000, select the first of the month
* Click *OK*

This is better, but it’s a bit too colorful. Also, when you click one line, you will notice that the others fade but don’t change to gray. We want to have a clearer selection. To do this, we need a little workaround. We will create a parameter that contains the selected city, and then use a calculated field to compare the city name of each data point to that. If it matches, it will turn blue, if not it will be gray.

* Right-click in the data pane on the left and select *Create Parameter…* from the menu
* Give it a name like *City to Highlight*
* For the data type, choose *String*
* At the bottom for *Allowable values*, pick *List*
* In the bottom of the window, use the *Add from Field* button to select the *City* field (this populates the possible values from all the different city names in the *City* column of our data)
* Select OK to close the window
* The new parameter shows up at the bottom of the data pane on the left
* Click it to show the menu and select *Show Parameter Control*

We now have a little dropdown box in the top right of the view, where we can select cities. That’s a start, but it doesn’t change the view yet when we select one.

* Right-click in the data pane and select *Create Calculated Field…*
* Name it something like Highlight
* Type the equation [City] == [City to Highlight]
  + Words between square brackets denote variables like fields, parameters, etc.
  + The field on the left is the name of the column in the data
  + The field on the right is the parameter we created above
  + The == operator is the comparison
* Click OK to close the formula editor
* Drag the newly created Highlight field from the data pane onto color
* If you only see two lines, drag *City* onto *Detail*. This tells Tableau how to break down the data. Since we only have two possible values from the Highlight field (*true* and *false*), that groups all the cities that aren’t selected together, which we don’t want.
* Click on the little triangle in the top right part of the color legend and select *Edit Colors…* from the menu
* Select a dark blue for the value *True* and a light gray for *False*
* In the color legend window, drag the blue color to the top if it isn’t already. That draws blue lines on top of gray lines, which is what we want.
* Now drag *Highlight* onto *Size* as well
* Click on the little triangle in the top right of the size legend and select *Edit Sizes…*
* Check the *Reversed* box to make the selected value heavier, rather than the unselected ones
* Adjust the size range with the slider and click Apply until you like the way it looks, then click OK to close the window.

This is the visual design part. Now for how to adjust where we calculate the index from.

* Create a new parameter called *Index Date* and set its data type to *Date*
* For the allowable values, select *Range* and set 1/1/2000 as the minimum and 7/1/2015 as the maximum
* Click OK, the click on the parameter in the data pane and select *Show Parameter Control*
* Click in the data pane to create another calculated field and name it *Index Relative*
* Enter the following expression: SUM([Index]) / LOOKUP(SUM([Index]), FIRST() + DATEDIFF('month', #1/1/2000#, [Index Date])) \* 100
  + This consists of two parts: SUM([Index]) / the Index at the location we want to do things relative to, and then multiply by 100
  + The second part figures out where the index for a particular data, specified by our parameter *Index Date*, can be found
  + The innermost part, DATEDIFF('month', #1/1/2000#, [Index Date]), calculates the difference between the date specified by the *Index Date* parameter and January 1, 2000 (the #1/1/2000# syntax is used to specify date constants).
  + We’re doing this for each city separately, so we need to keep track of where in the table the first value for that city is, which is what the FIRST() function tells us. In the table on the right, the first row in each section is pointed to by that function.
  + The LOOKUP function then pulls a particular value of Index from the location we specified so we can divide the value we’re interested in by it.
* Click OK to close the formula editor
* Drag the newly created field into the *Rows* shelf and drop it on the *Index* field, so that it gets replaced.

Now when you change the date selected in the Index Date control, the visualization will update and show the recalculated data.

Finally, we want to see the index date on the visualization and also show a line at 100 as a reference.

* At the top of the data pane on the left, click on the *Analytics* tab
* Drag a *Reference Line* from the *Custom* section into the view and drop it there
* In the window that appears, click on the drop-down next to *Value* and pick the *Index Date* parameter
* Next to the *Label*, pick *Value* (or *Custom* and enter text like *Index Date*)
* Click OK. This shows a vertical line at the index date (try changing it to see what happens)
* Drag a *Constant Line* from the *Summarize* section and drop it on the *Index Relative* field that appears.
* In the text window that pops up, enter 100 to set the line to the 100 reference value
* Next to *Label*, pick *None*
* This now shows a horizontal line at 100

To keep the view stable as you make changes, double-click on the vertical axis and select a *Range* of *Fixed*. Set the start to 30 and the end to 300. Delete the title as well for a cleaner look. Now as you change the index date, the values are always comparable because the visualization does not try to rescale to the data.