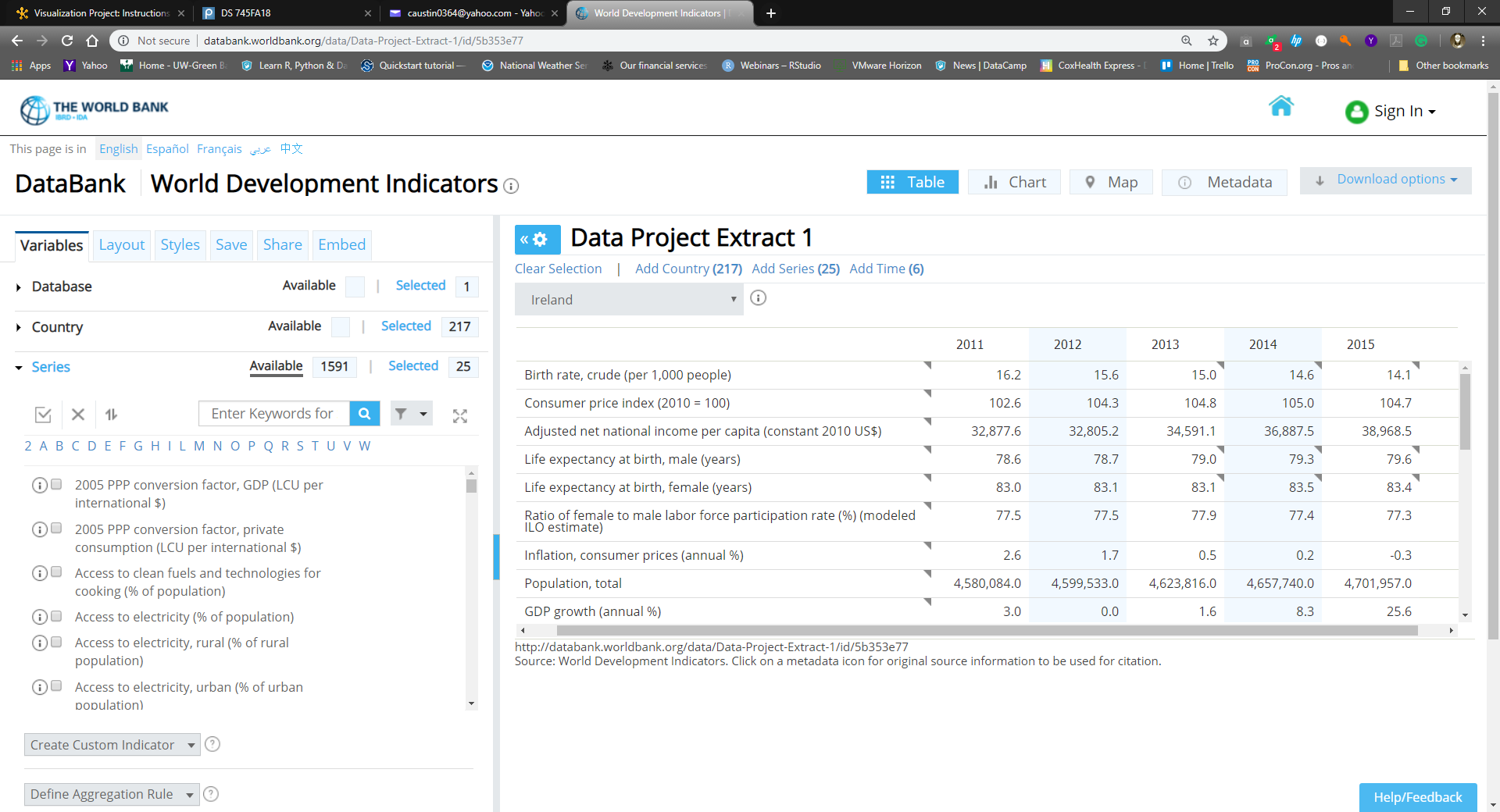
Purpose of Visualizations – Story:

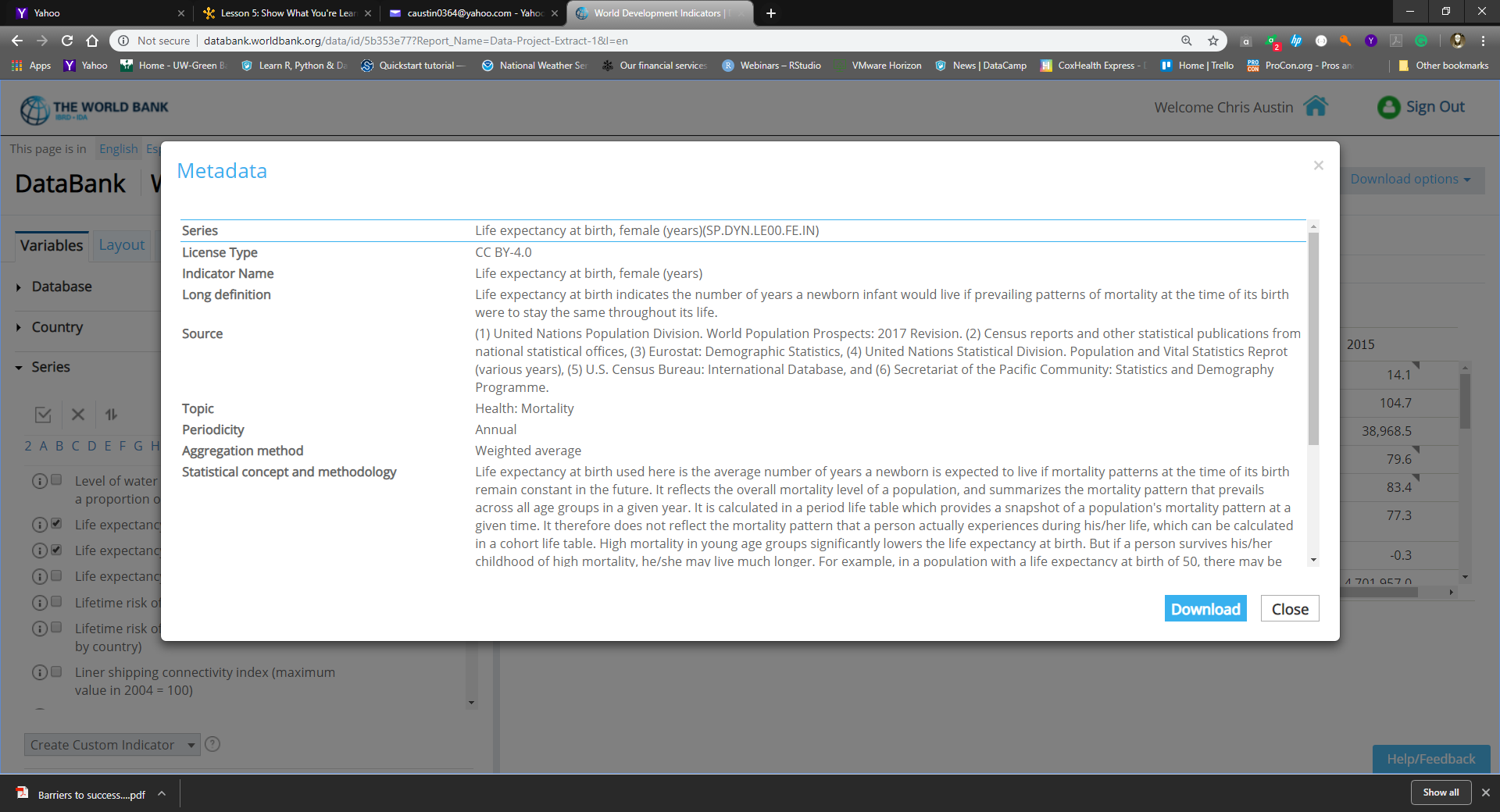
The goal of this visualization is to illustrate that women’s health - as measured by life expectancy - is improving over time on a broad global level, regardless of specific countries’ birth rates or poverty levels – as measured by Gross Domestic Product (GDP).

Purpose of Visualizations - Background on dataset:

The dataset was pulled from the World Bank’s DataBank[[1]](#footnote-1). Anyone is free to extract data from their query engine, which does an excellent job stepping you through the process. You also have the ability to save your queries for public view if you have a user profile. Your Twitter, Facebook, or LinkedIn profile is sufficient to log in with.



Looking for variables I felt would be related to the question at hand, there are nearly 1600 series to choose from in just this one of dozens of databases. Each measurement has an info icon so that one can identify who the World Bank is citing as the author, full definition, and circumstances of the measurement.



I was able to create an extract file for easy download[[2]](#footnote-2). Readers of this document can plug in the link and verify the data selected, or download it themselves if needed. I only had to delete a few lines at the bottom of the CSV file related to disclaimers before I used it.

What you’ll find at the link is a superset of the dataset I eventually ended up using. With the amount of data available, I had to build a base set of data for exploratory purposes prior to committing time to the visualization.

Choice of variables:

After the exploratory investigation of the dataset was complete, there were five variables chosen for the final version of the visualization. They align well with the story; living conditions and the health of women are continuing to globally improve. The dataset has no missing entries apart from occasionally unavailable GDP figures. Those instances were plotted differently so that the rest of the data’s context could be used, ensuring that no data goes to waste.

217 countries and protectorates were selected.

Time periods chosen were 2010 – 2015.

Birth rates per thousand people.

GDP per capita, in 2010 US$. This allowed for an ‘apples-to-apples’ economic baseline comparison.

Female life expectancy in years.

The axes chosen were birth rates per thousand (x) and female life expectancy (y). The act of childbirth is historically one of the most medically taxing events - if not the primary one - in a woman’s lifetime. As such, it seems appropriate to pit the two against each other to visually gauge their correlation.

Choice of tools, methodology, and design principles:

Being somewhat familiar with R’s *ggplot*, I chose to go that route. It’s a high-quality package based on Hadley Wickham’s interpretation of Leland Wilkinson’s *The Grammar of Graphics*. I wanted to work towards Tufte’s ideal of maximizing data ink and minimizing erasable ink, attaining a high data-ink ratio. I feel the final visualization delivers on that goal.

The RColorBrewer package was used to provide meaningful color palettes. I created some code to take the GDP values and bucket them into ranges so they could easily be color-coded. I saved converting the NAs for last in order to give them a special color outside of the range seen in the brewer.pal() call.

colorCount1 = 7

cut1 <- cut\_number(wbDataSpread$GDPPerCap,colorCount1)

gdpColors1 <- as.numeric(cut1)

# Set all NA’s to colorCount1 + 1 and add column to dataframe

wbDataSpread$gdpColors1 <- as.factor(ifelse(is.na(gdpColors1), colorCount1+1, gdpColors1))

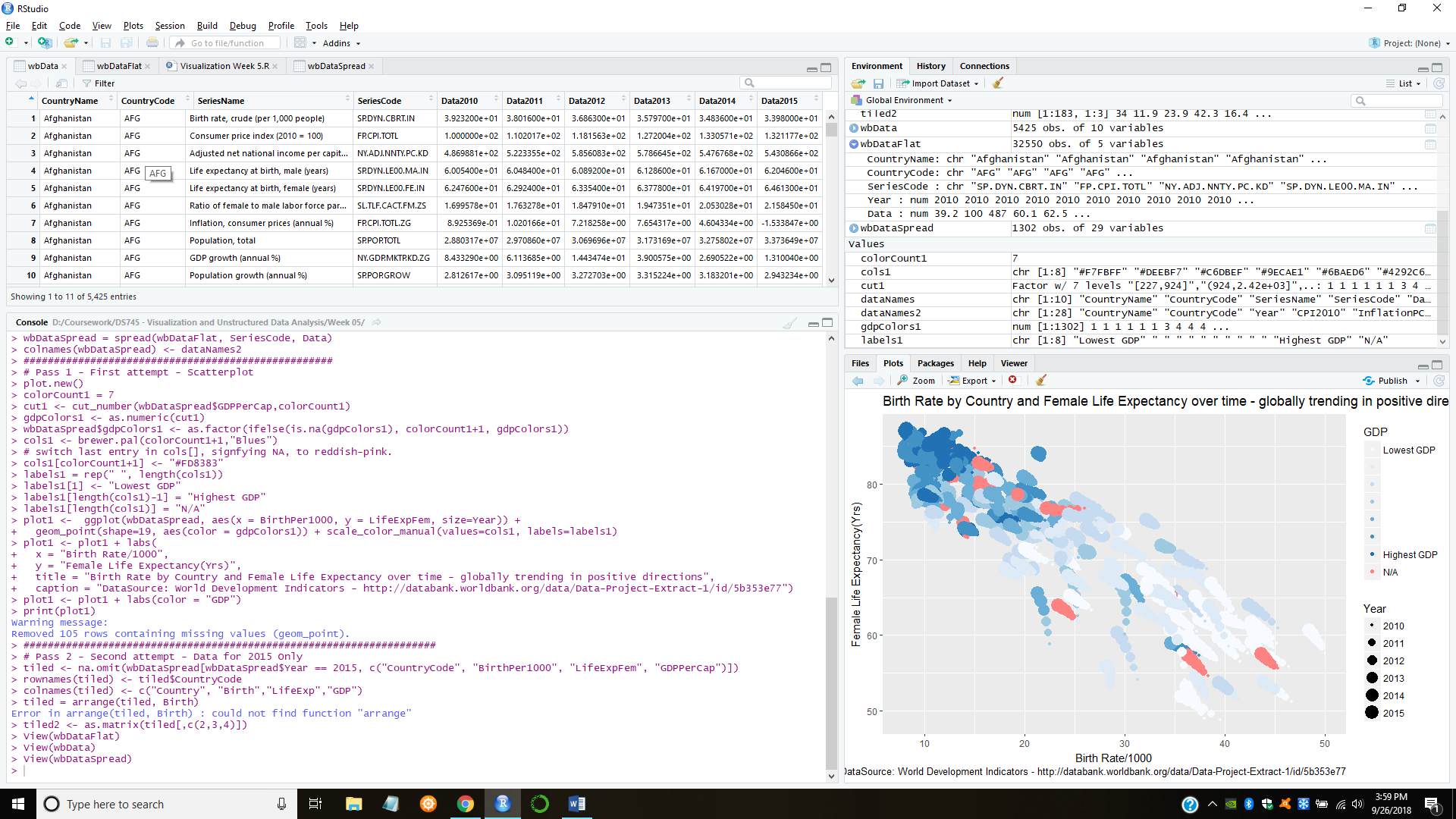
cols1 <- brewer.pal(colorCount1+1,"Blues")

# switch last entry in cols[], signfying NA.

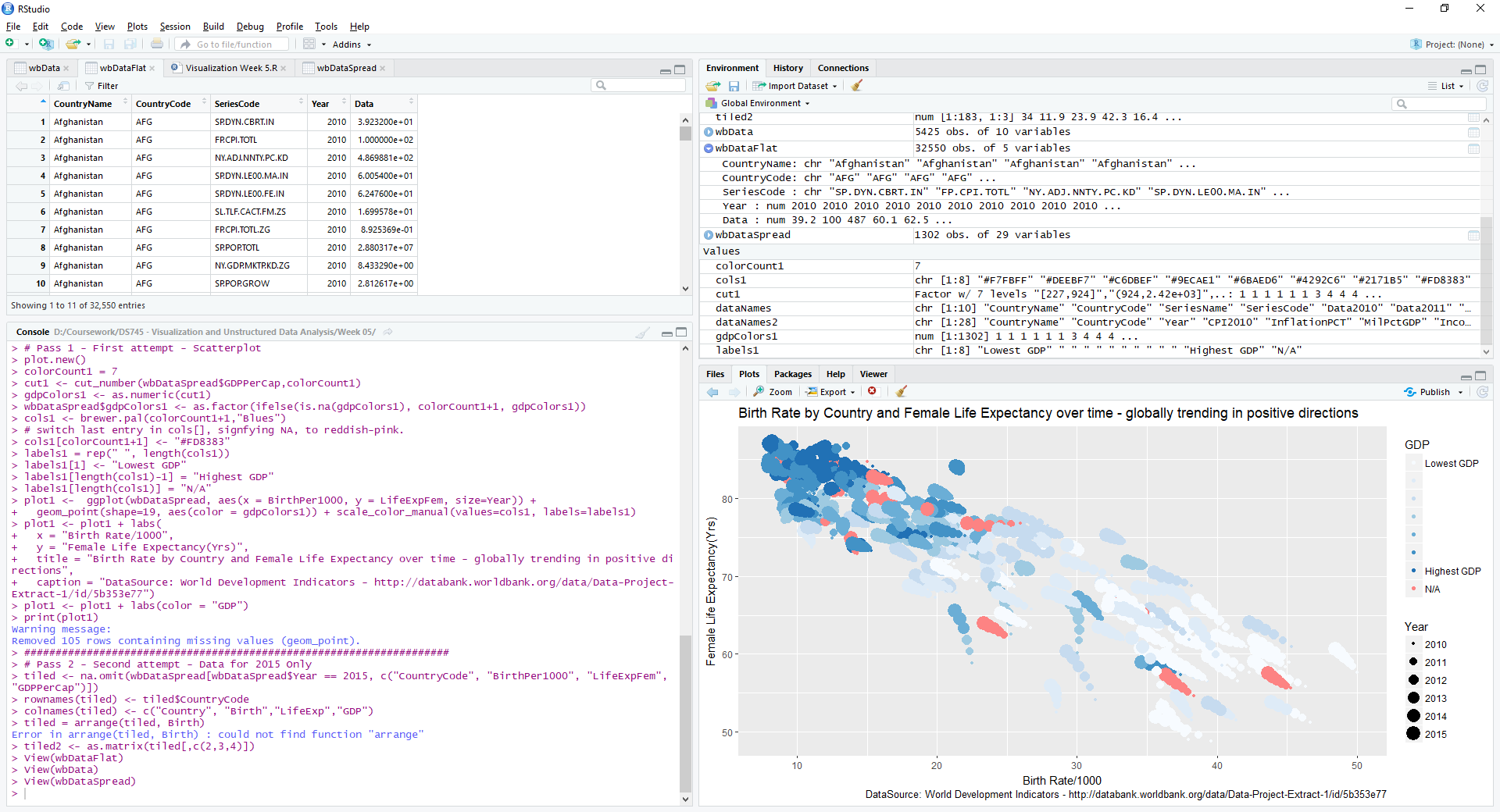
cols1[colorCount1+1] <- "#FD8383"

Prior to employing ggplot, the data had to be munged in three passes. Methods out of Wickham’s “tidyverse” were utilized: tidyr:spread() to spread row-centric data out into columns for each country and year, and later dplyr:arrange() was used to sort dataframes by column for the second visualization.

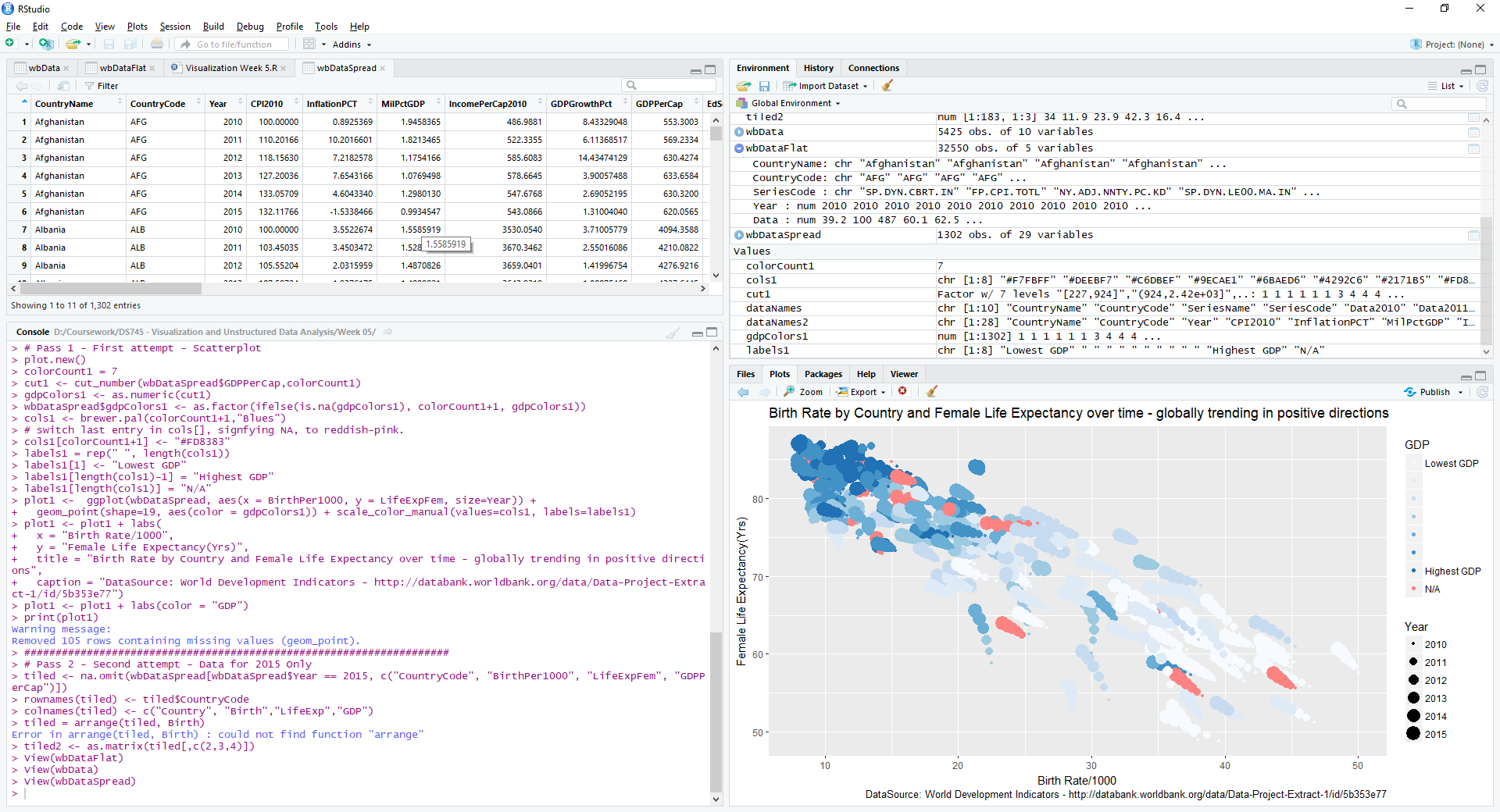
Raw Download from World Bank



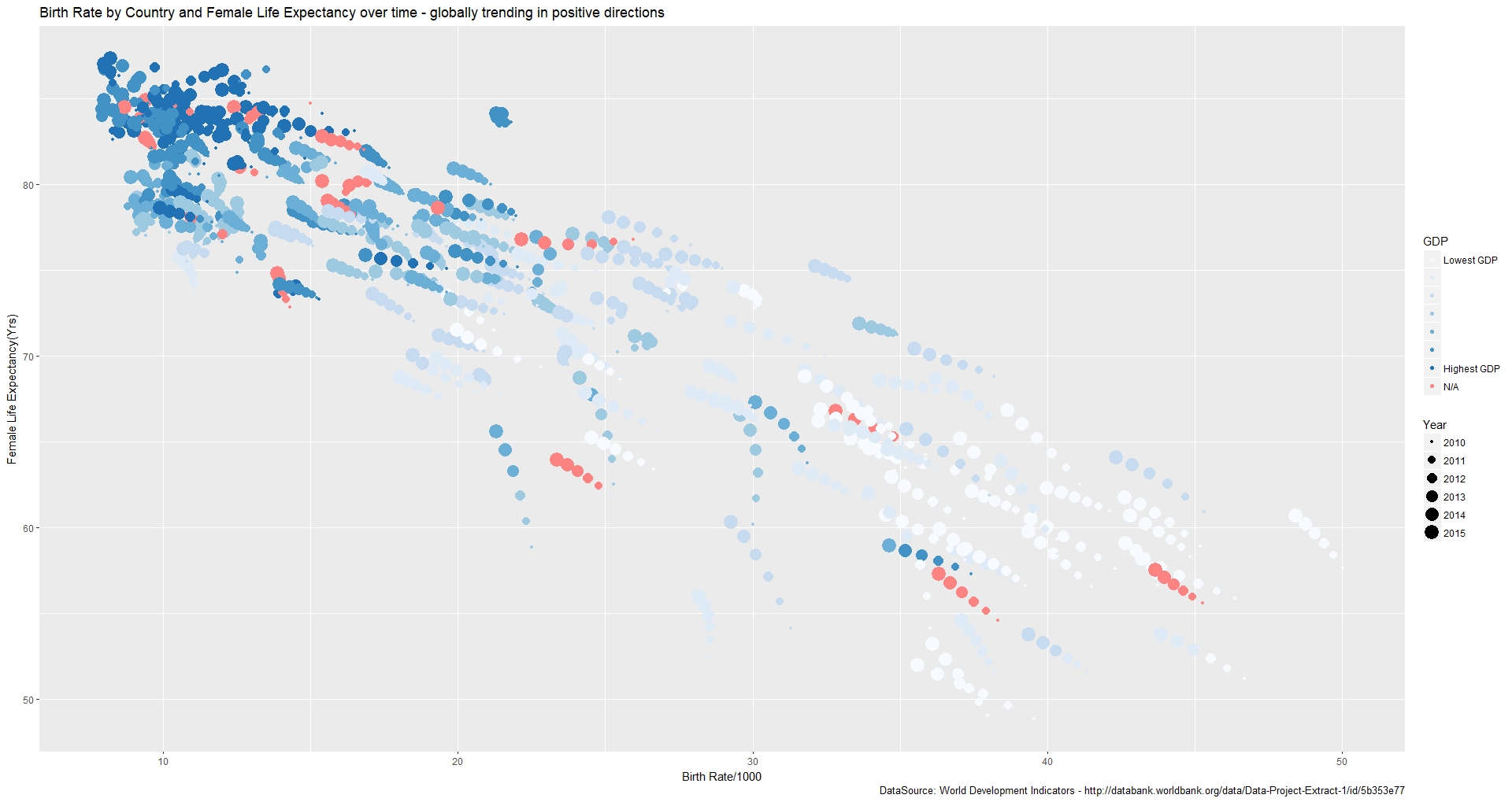
After creating one dataframe/year in order to cbind() them together



After using spread() to change rows to columns



Explanation of Viz 1 – Scatter Plot, First Pass:



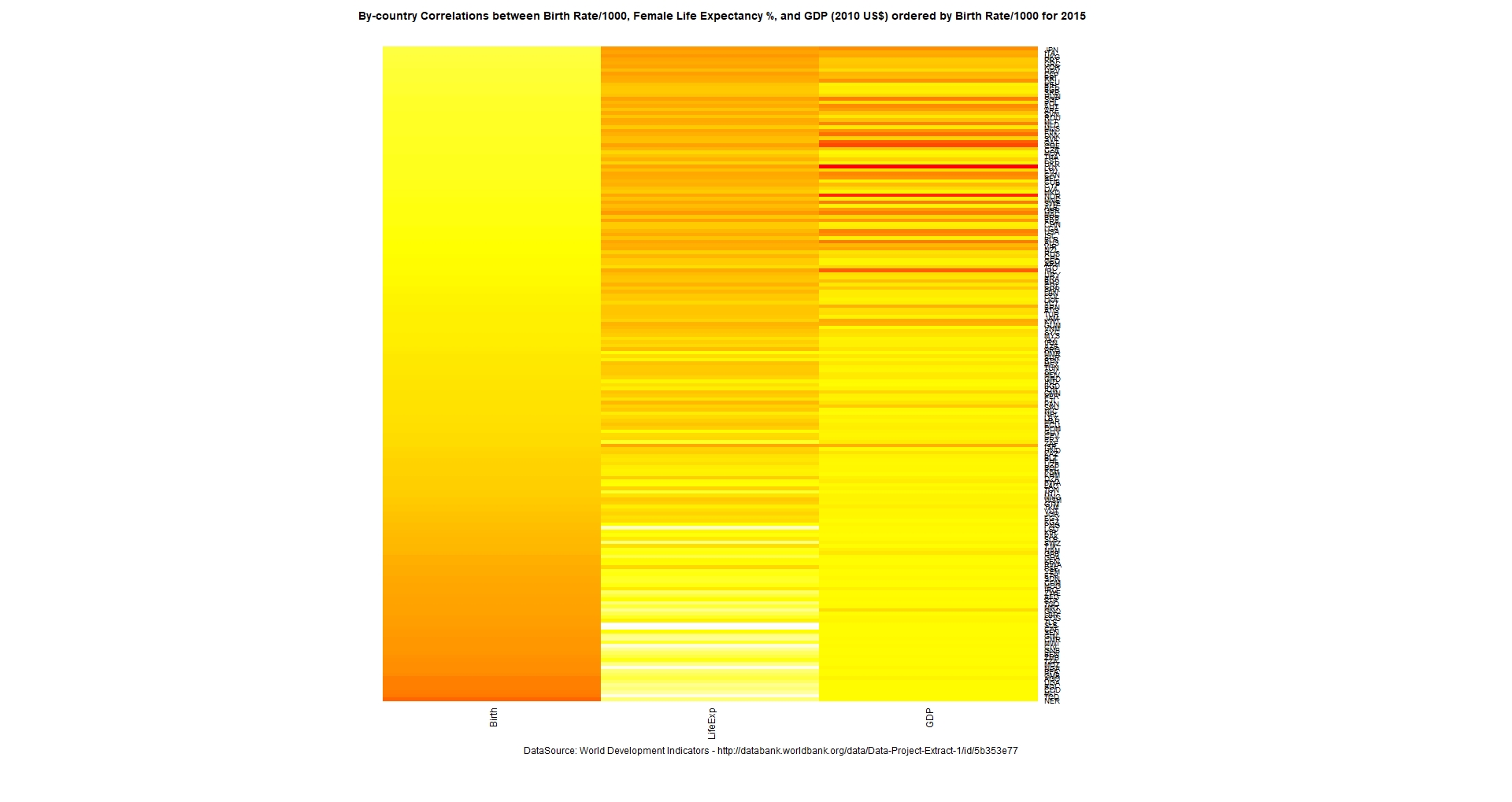
Pros:

* Having the year column determine the size of the data point makes for a good simulation of motion / improvement. It certainly allows the viewer to see general patterns are consistent for most countries.
* With the gridlines being white against the gray background, they don’t interfere with the viewer’s interpretation of – and reasoning about – the visualization.
* Viewers are drawn to making interpretations of what the various slopes in data patterns signify. Some show equal rates of decrease for birth rates and increases in life expectancy. Other patterns with a steeper slope show mostly decreases in birth rate. Additional groups show nearly flat slopes, signifying decreases in birth rate without offsetting increases in life expectancy.

Cons:

* The color for N/A is a little too distracting. Need to tone it down in relation to the rest of the palette.
* The light colors of RColorBrewer’s “Blues” palette are a bit too pale. They only work because of the gray background.
* Despite the gray background being a positive in drawing out the lighter elements as noted earlier, it still runs a bit counter to Tufte’s concept of minimizing erasable ink.
* Specific countries are not identified. The grouping of data in the top left corner made that impossible.
* The title communicates the primary theme of the story, but can be expanded upon.

Explanation of Viz 2 – Heatmap visualization:



Pros:

* heatmap() allows you to throw a matrix at it, presenting as many series as you care to show together.
* When ordered by Birth Rate/1000 for 2015(only), the heatmap does a good job at using the color scheme to show the rate of change in values between the three series for all countries. This prompts the viewer to conclude that lower birth rates strongly correlate to higher female life expectancies and to higher GDP.
* The country labels are all visible. Perhaps those would show better as a separate data grid if one was curious as to where a specific country placed on the chart.

Cons:

* heatmap() has a limited number of options geared towards better presentation. ggplot’s geom\_tile() only looks for vectors in dataframes for x, y, and fill parameters.
* Trying to plot more than one year’s worth of data didn’t help. An earlier attempt to plot 2015 data along with the 2010-2014 data took away from the message of correlated variables that the visualization is trying to elaborate upon. This keeps the viewer from perceiving the time-value element of the story.
* Lots of messing with R’s plotting defaults were necessary to eliminate title and subtitle overlay.
* The country labels are too small to make out with the quantity of data presented.

Explanation of Viz 3 (Final) – Enhanced Scatterplot:



Pros:

* The grid has been removed from the first iteration. Axes as well. The sparseness of the visualization helps bring out the lighter colors.
* The legend has also been removed to maximize data space. It has been replaced by labels that point out examples of screen element meanings (missing data, changes in color due to GDP levels, and changes in size due to time progression). Furthermore, the color for the N/A GDP values has been lightened from the first iteration so that it doesn’t clash with the rest of the visualization and take attention away from the rest of the data.
* There looks to be very little - if anything - that can be removed without muddying the overall message: convincing the visualization’s audience that women’s health around the world is generally improving in conjunction with other factors. While time is admittedly not the best of explanatory variables, GDP was added in order to pair a good explanatory variable with a poorer one and shore up its effect on the viewer.
* A subtitle has been added to point the viewer to a secondary observation – what’s next for the countries at the top left of the chart? There’s a bottleneck not just at a nearly-impossible zero birthrate, but in maximum life expectancy due to lifestyle, genetics, and medical care. The only way forward (given the impossibility of a negative birth rate) is a vertical move up the y-axis, awaiting the next medical advance affecting longevity.
* Note the flip of the color intensity of GDP values, higher being lighter and lower being darker. It probably seems a bit counter-intuitive (and counter to best practices as covered in class) to do that, but our reading has also mentioned that there are times to bend those rules. I did so here because a) I wanted to emphasize improvements women’s health (again, measured by longevity) in more impoverished countries, and b) the bottleneck at the top left. If the bottleneck remains heavily-colored as in the first iteration, it draws your attention in that direction due to an “ink overload”. That’s why, in the subtitle, I tried to re-cast the metric **as a decrease in poverty** rather than an increase in GDP.

Cons:

* Limited labeling of country names: An attempt to add all country names resulted in a detraction from the overall message. It would have been good to put all the country-name labels in there, but data is a bit bunched together at the top-left. In lieu of that, I tried to add a few countries to provide either a) references at the extremes of a plot or b) outliers. I was reminded of a visualization in the NY Times seen in DS735 that made me think of what the chart would be like if interactive; individual countries could be chosen, continents could be highlighted and other stories could be told:
  + <https://www.nytimes.com/interactive/2014/06/05/upshot/how-the-recession-reshaped-the-economy-in-255-charts.html>
* Still a bit of a gathering of data at the top left of the chart of the low birth rate/high life expectancy countries. It would have been nice to separate them but I think that would have taken another chart dedicated to that section and a different scale for the axes. But, that’s one of the findings: We’re trending to more common life expectancies and birth rates.

1. http://databank.worldbank.org/data/home [↑](#footnote-ref-1)
2. http://databank.worldbank.org/data/Data-Project-Extract-1/id/5b353e77 [↑](#footnote-ref-2)