Alex Johnson
SI 601

Visualizing World Maternal and Child Health Stats and US Aid

Motivation

Health is a major issue that exists throughout the world and where the United States dishes out aid to developing countries can be quantified in many ways. In particular, for this project, I wanted to make a point to focus on maternal and child health. It was my intention to see if there was a relation or correlation between the amounts of US financial aid given to countries throughout the world, developing or otherwise, and compare it with figures/data maternal and child health. It was my hope that through analysis of the data available, some kind of visualization could be created that would point to trends and commonalities with the datasets that would be eventually parsed. The main goal and main question was whether countries that received more US aid were positively impacted in conducting reliable and effective programs for maternal and child health, if that was even an answerable or quantifiable question. The World Bank dataset series called GenderStats and the API from foreignassistance,gov were my best bets for providing the type of answers I hoped to acquire. My hope was that the analysis would answer just how big a role that aid plays in initiating valuable health programs for women and children.

Data Sources

For my analysis, I used two datasets: a csv file called GenderStats from the databank at the World Bank and an API call to foreignassistance.gov. The GenderStats dataset is located at http://data.worldbank.org/data-catalog/gender-statistics, and was opened and read as comma separated values. The GenderStats dataset (totaling 19.8 MB) covered the years 1960-2013, and had 346 separate statistics on 261 different countries throughout the world. The important variables within the dataset were a matter of what statistics I felt needed to be kept for the analysis, and what year to focus on. For this project, I focused on 7 statistics that were relevant to my main motivation: ‘“Number of maternal deaths”, "Mortality rate, female child (per 1,000 female children age one)", "Maternal leave benefits (% of wages paid in covered period)", "Pregnant women receiving prenatal care (%)", "Contraceptive prevalence (% of women ages 15-49)", "Met need for contraception (% of married women ages 15-49)" and "Mortality rate, infant (per 1,000 live births)”. All 7 I felt would help me visualize the story in the way that I wanted and to give me valuable insight on these health issues. After looking at the available data, it became apparent that data for maternal and child health is underrepresented worldwide as a whole, especially pre-2010, and in order to maximize results (there also being a lack of sufficient data for 2013), I chose to focus on 2011 which was recent enough, and had the most data for the statistics I intended to study.

The API I used from foreignassistance.gov is located at http://www.foreignassistance.gov/web/Developer.aspx. It was returned as a JSON string using the urrllib2 module and loaded with the JSON module. I was able to filter it by sector (maternal and child health), year (2011 & 2012), and the type of transaction (spent). I decided to return two JSON strings for both 2011 (65 records total), and 2012 (69 records total) in order to create a more complete picture on the impact of US Aid. It was my intention to take the statistics from the first dataset and see how much of a rise in US Aid was given as a result or in response. The API gave me the option of searching for planned, obligated, spent and total transactions, however it was my decision to filter spent monetary values only, as it represented the actualities rather than the intended. The JSON string dictionary key/value pairs returned 'BenefitingLocation', 'FiscalYear', 'Sector', 'Amount' and 'AgencyName'. Only 'Benefiting Location' and 'Amount' were necessary to keep, as 'Sector' and 'Fiscal Year' were implied by the filter of the API calls and I was looking at all US Aid spending, so 'Agency Name' didn't matter. Of the 65 records for 2011 and 69 records for 2012, I ended up only keeping 54 records from both as I started coding; some records like 'Benefiting Location'; 'Central Asia Regional Office (USAID)' didn't match or correlate with any of the 261 countries listed in the comma separated value dataset from the World Bank.

Data Manipulation Methods

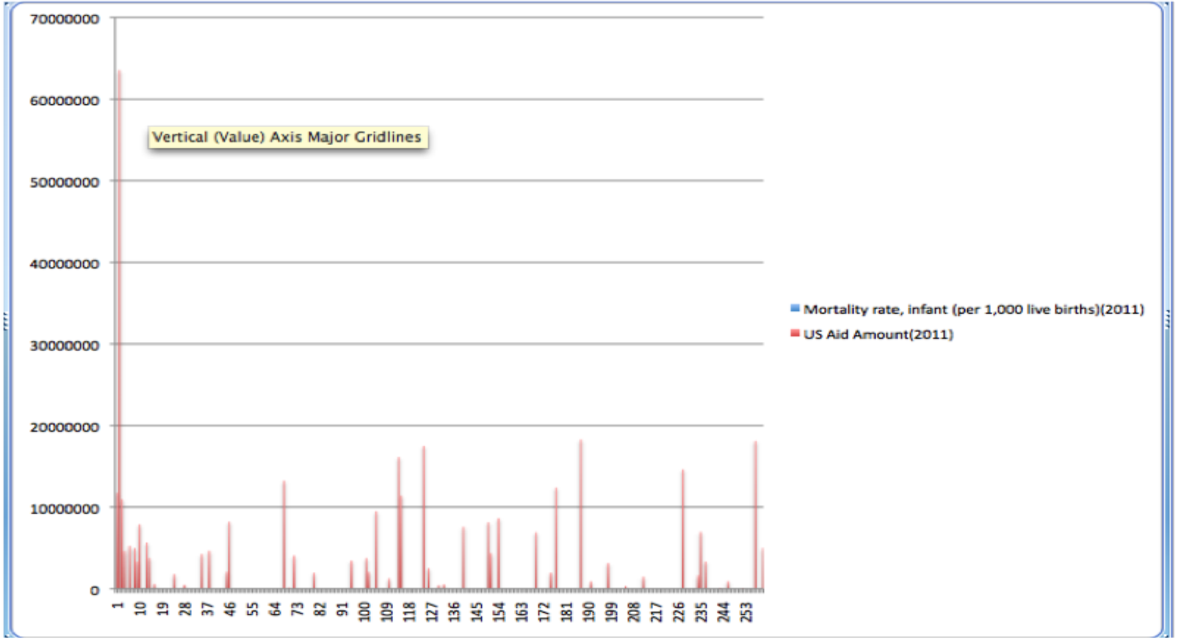
The main task for the GenderStats dataset was to filter each line based on the intended statistic and determine what line index it referred to. Once I got those parsed for those lines, I could easily grab the other data I wanted based on what their index was in the line (since I was only keeping lines with the 7 statistics, no other rows mattered). I was able to append each data point to separate lists based on what condition (statistic name) it met, and since I would need one instance of the country name, I appended to a list under only one if/elif condition (aka one statistic - it didn't really matter which one, every country had the same statistics). Because there was so much missing data, I needed to create a condition that would allow me to put values as null if my code met a blank cell. For instance if line[2] for "Number of maternal deaths" and line[55] for 2011 was blank, then line[55] = 'Null', and would be appended to the maternal death list, and so on and so forth. The lists from the first dataset were then zipped into a giant list with data that was eventually captured from the API. Each item of that analysis was then written out as a comma separated value. To manipulate the JSON string from the API into the eventual list that would be zipped, I needed to create temporary dictionaries using country; aidmount as key/value pairs, then append that dictionary into a list for the max length of each JSON string. Getting the data for the API was a lot trickier than the first dataset, because some countries from the country list for the first dataset didn't exist, meaning they did not have US Aid data available. Additionally, because there were multiple aid amounts for some countries, I had to iterate through the list of dictionaries and create another dictionary for country/amounts again, adding the totals for similar keys as values. Finally I had to take in account missing aid amounts and had to match country names from the last dictionary with items from the initial country list, inserting null values where no data existed: finally appending those aid amounts into a list. I then was able to zip up all lists of countries and statistics from the first data set and the aid amounts from both JSON strings into a giant list as discussed previously, which I then wrote out to a comma separated value file. No utf-8 encoding/decoding steps were necessary for processing the analysis.

To put it more succinctly, the workflow for the source code to get the final analysis is as follows: 1) open the GenderStats csv and return the JSON strings from the API 2) extract the country and aid amounts from each max len JSON string into a dictionary and append those key values to a list 3) iterate through that list and creating another dictionary that sums values (aid amounts) for similar keys (countries) 4) iterate through each line of the GenderStats csv, using if/elif statements to find statistics, and if/else statements to deal with missing data, ultimately appending each statistic result to its own list 5) iterating through the country list to match keys from the most recent dictionary for the API, appending corresponding values or 'Null' values to a list for aid amounts 6) use the zip function to create a giant list 7) write those items out to a csv file.

I found there to be two main challenges: how to get the right aid amounts to correspond with the right countries, and how to best represent the data when there wasn't a lot of data for the years I wanted to begin with. The syntax for creating the right list of aid amounts took a lot of processing, as has been discussed previously and processing that JSON string took the most amount of time. The second problem had more to do with what data was available in the GenderStats dataset, which affected how I was able to ultimately analyze the two data sets that could best give some kind of answer or solution to my original question. The main problem with my findings was that a majority of the data points from the GenderStats dataset were missing - meaning that those countries never reported the data, or the World Bank was unable to determine accurate rates and ratios for that data. I believe it refers to a larger problem, that maternal and child health is an underrepresented 'demographic' in the larger sphere of public health. Given more time, or more data resources, I believe that a more accurate depiction of the impact of US Aid on maternal and child health can be established, but more efficient data needs to be reported or discovered. Additionally, the dataset lacked most information for the year 2013, which is a problem because the most recent impact of US Aid could not be established as a result - I instead had to focus on 2011 data, and tried to correlate that with US aid figures from both 2011 and 2012. This meant that only a retrospective look at past trends could be discovered, and present solutions and trends only predicted...since there was insufficient data to work with. Another dilemma was whether to focus on the larger dataset of 'transactions' from the API vs. 'spent'. Transactions meant amounts that were planned, obligated, and spent as has been discussed. I chose to focus on actual spending figures for a more accurate picture of impact. While a complete analysis can not be completed based on the data provided, I do think it points to the need for this data to exist so that sources that finance US aid can accurately determine where and how much aid should be sent for maternal and child health issues.

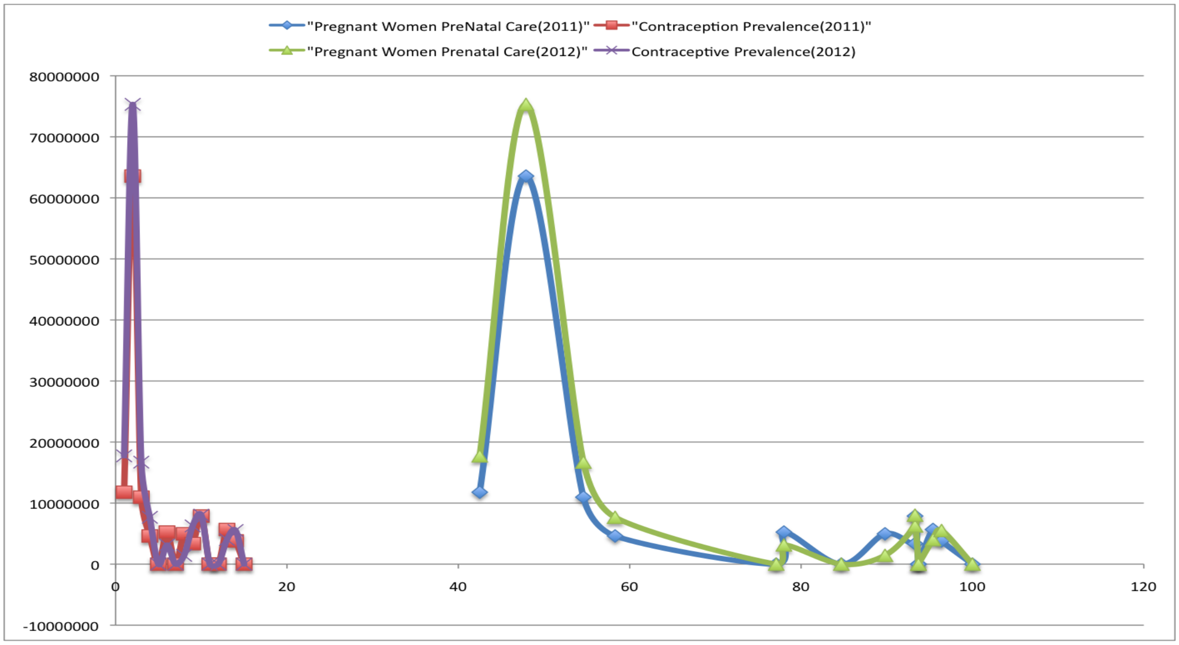
Analysis and Visualization

The final step in my workflow code allowed me to produce an analysis in comma separated values, where I wrote out the items of the zipped list using the csv module. Given that only 54 countries had aid data, I had to ignore 207 countries to make any sort of comparison. Additionally, the only column of statistics that was full enough to make a correlation for anything health related was infant mortality rate. I decided to graph that statistic with both sets of aid data for 2011/2012 to see if I could see a trend of positive increase. That chart is below and created with excel:

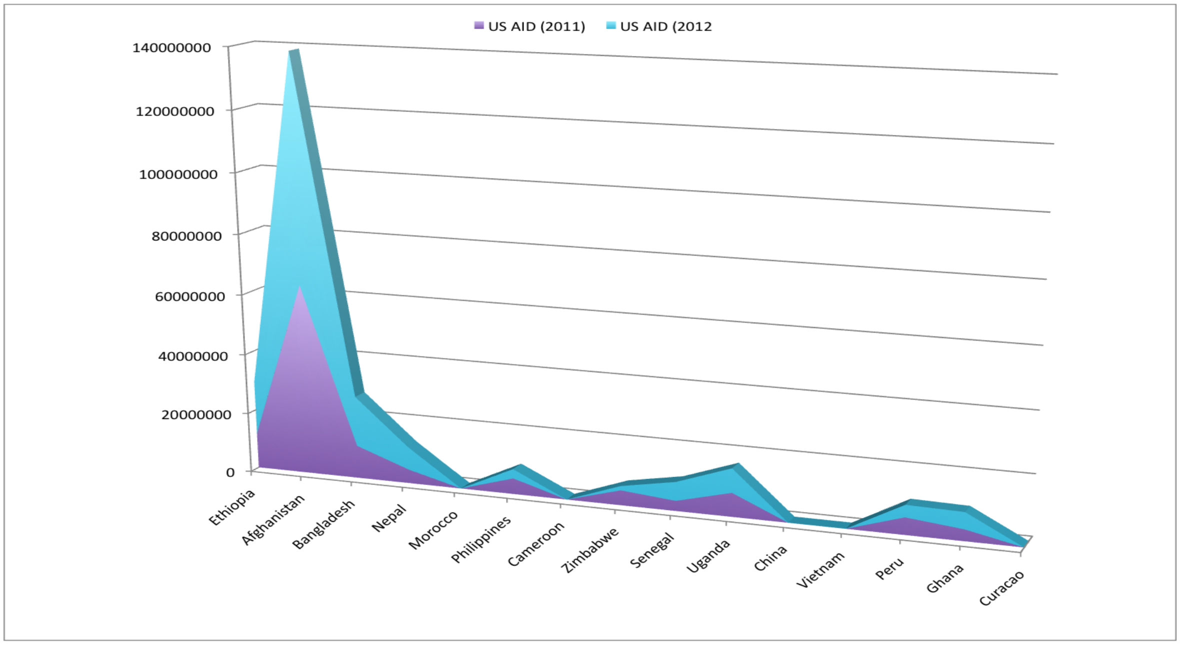


For infant mortality, interestingly, the data doesn't seem to support my original hypothesis, however given the insufficient data and null values, I am unsure if an accurate depiction can be determined. Additionally of interest: over 60 million US dollars has gone to Afghanistan (if you were to compare that chart amount with the csv analysis country column). This probably is more politically motivated because of the years long US war and subsequent destruction. For the other statistics, I decided to sort those columns in order to grab the data with only non-null values.

The X, Y scatter plot graph below, made in excel, represents the 2011 data for pregnant women receiving prenatal care and the prevalence of contraception. I compared the same data with aid amounts for 2011 and 2012 in order to see if there was a significant change in how US Aid resources were divvied up. There appears to be a minor correlation in rate of care and percentage of contraception use with aid totals in this case.



The countries, while limited, correspond with the chart below it, which is an overview of aid data for both years. More aid was given out in 2012 than the previous year, and the big spike in that chart, like infant mortality rate, represents Afghanistan.   It would be interesting to know if the US Aid totals were given out not as just a response to maternal and child health needs, but factors like war, poverty and natural disaster. It would appear that more analysis will need to be done in order to develop a more clear picture of reasons why one country would receive aid over the other.



Conclusion

Due to insufficient data, I was unable to fully quantify or answer my original question and motivation. I do believe that value exists in the two datasets I used, and that the given other parameters, a sufficient analysis can be conducted. I believe the GenderStats dataset is continually updated, so perhaps in the future, the data scientists and humanitarians at the World Bank will able to discover data that fills the blanks. Perhaps this is more an indication of transparency, both political and institutional, but perhaps a clearer picture will be uncovered, through more reporting and more accountability.