EDAV-Project2

Ain't no sunshine

March 08, 2016

Introduction

For this report we will be diving into the global issue of Flood Events. The intial goal of the paper will be to investigate the datasets to better understand. The dataset being investigated involves geographical, spatial, and time parameters that creates an extremely large amount of information to be explored. We will work to break down this large dataset into something smaller and begin to focus on specific events within the flood dataset.

An nteractive google map with tagged flood locations (1985-present) is displayed below. This feature allows the user to zoom in and see exactly where a flood event occured. Clicking the tag shows the date of the event.

Global Flood Events: From Januaray 1st 1985 to December 23, 2015

Word Clouds

In order to understand the map data we begin to break it down by different categories and look at countries with the largest totals in each using wordclouds to investigate the different components of flooding events. The first wordcloud shows the 100 countries with the greatest number of events since 1985. The USA and China seem to be the two countries that were most affected, with the Philippines, Indonesia and India close behind. The second wordcloud highlights the attributes of the "detailed locations" column in the dataset and gives interesting insight into the types of areas that are commonly affected by flooding.

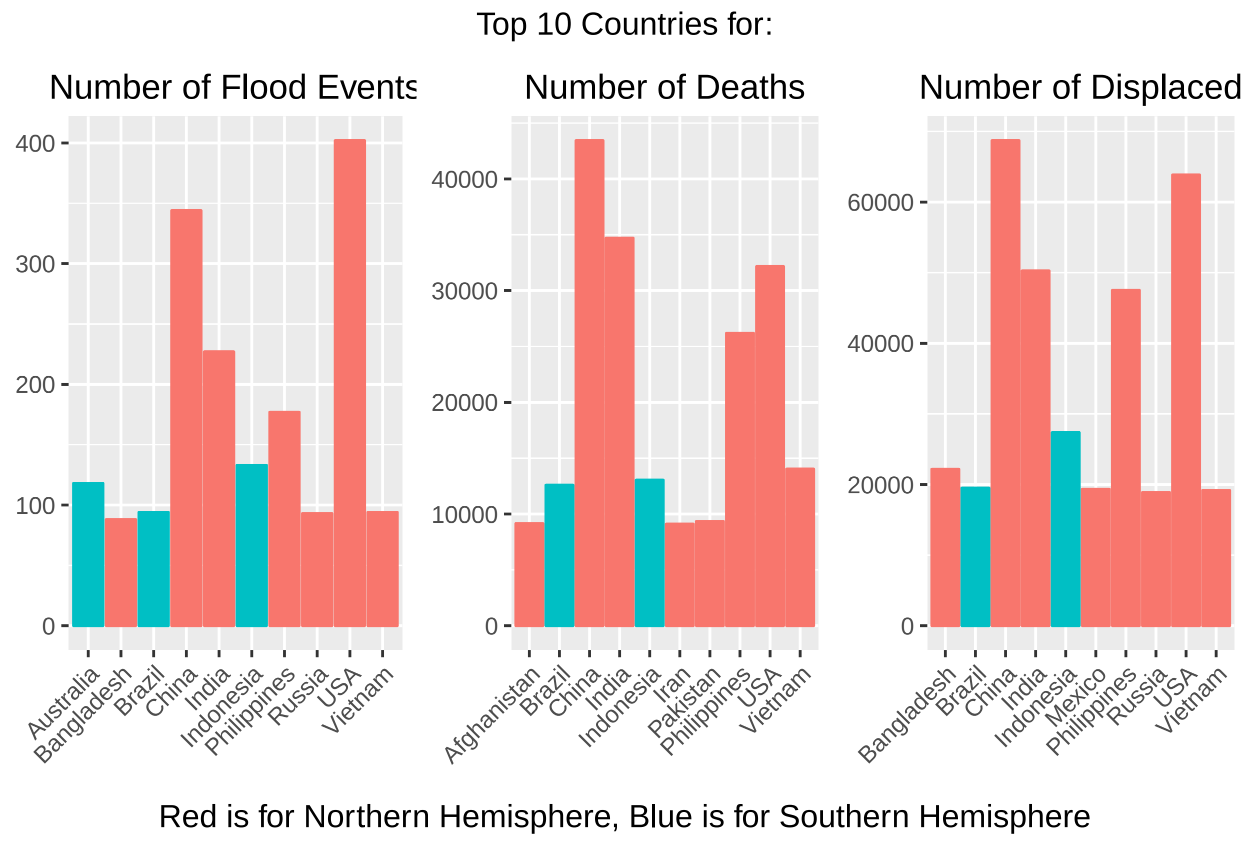
Countries Word Cloud   Locations Word Cloud

Word clouds were used to try and analyze the differences between the Northern Hemisphere (postive Latitude) and the Sourthern Hemisphere (negative Latitude). We wanted to investigate the column labeled "Notes and Comments", which included many news articles after cleaning the text the following word clouds were produced.

What we can see from these plots is an interesting split between the Northern and Southern hemispheres. "People", was largely contributed by the Southern Hemisphere and "Flooding" was more attributed to the Northern.

Top 10 Country Investigation

Now we will investigate the top 10 countries within different categoris of the flood data. To see if we can visualize any unique characteristics within the data. We will look at variables such as the number of floods, deaths by the floods and people displaced by the floods.



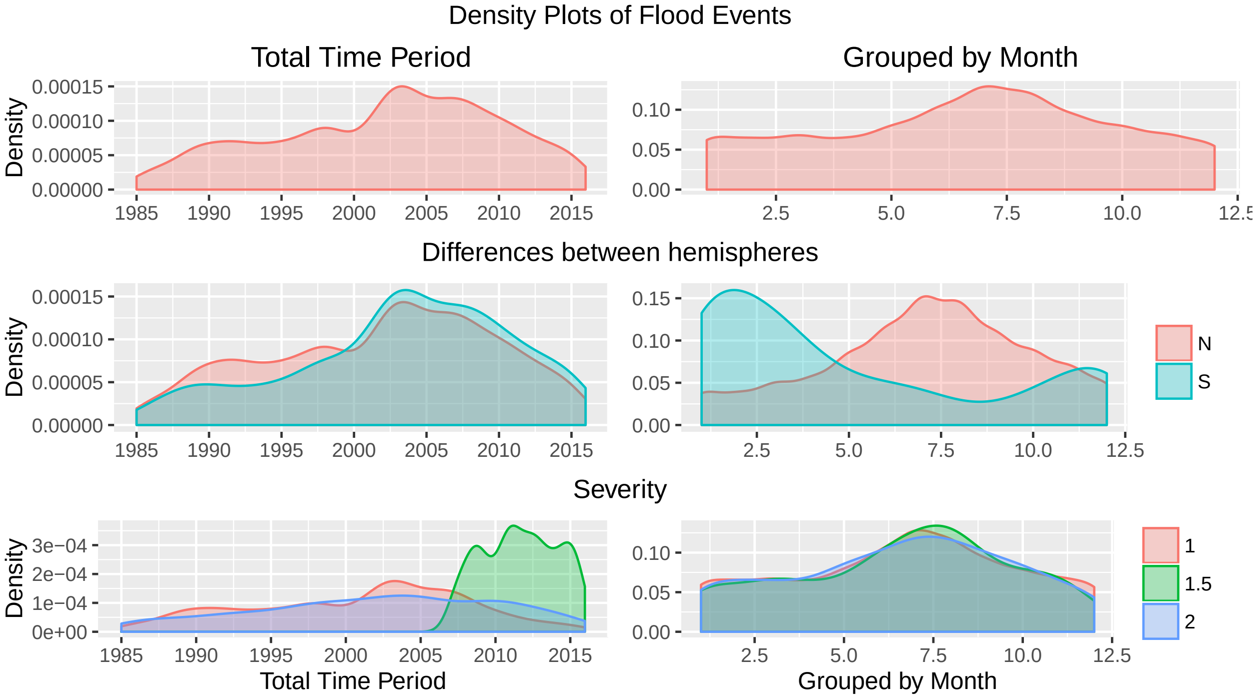
We can see that there are some insights within some of these plots. We can see that there are several countries that appear in all three plots. These are Brazil,China,Indea,Indonesia,Phillippines,USA and Vienam. Seeing some of the more devloped countries such as the USA seems surprising here because of how many flood events actually occur. There could be a difference in reporting here as well, such as the USA has been recording these events longer and therefore has a higher number of events.

Another interesting finding is that Iran, Pakistan and Afghanistan appear in the highest number of deaths but not in the highest number of displaced. Looking deeper into these categoies we find that they just mostlikely missed the top 10.

- Iran we can see that there were 9,107 deaths which also accompanied 10,364 displaced.
- Pakistan we can see that there were 9,346 deaths which also accompanied 14,611 displaced.
- Afghanistan we can see that there were 9,150 deaths which also accompanied 10,931 displaced.

So we can see a positive correlation occuring between the number of Flood Events, Deaths and Displaced.

We then continued looking into the differences between northern and southern hemispheres looking at density plots of varios attributes of the flood events.



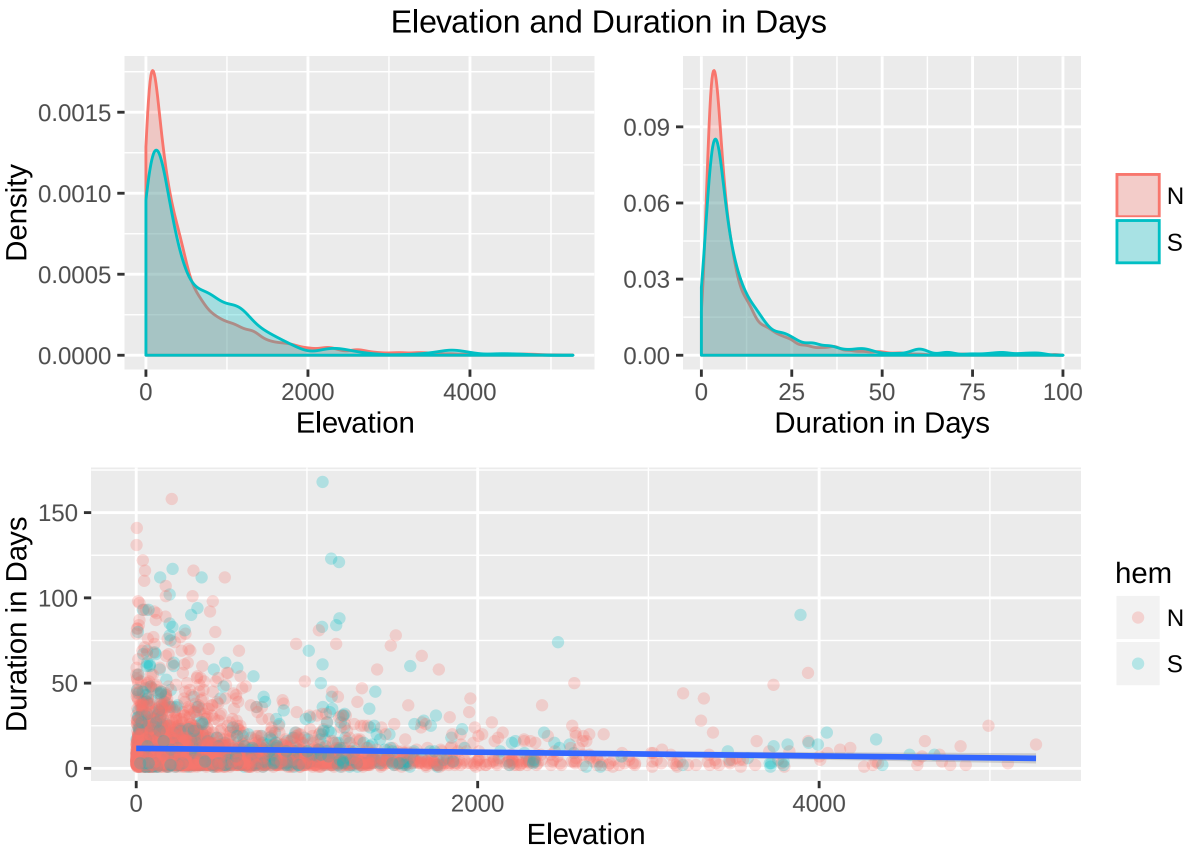
These density plots show us that the total number of flood events occured between the years of 2002 - 2003 by looking at the left column of plots "Total Time Period" Looking at "Differences between Hemispheres" we can actually see this may be mostly contributed by the Southern Hemisphere Countries begining to record their Flood Events along with the countries in the Northern Hemisphere. We can see that from 2002 on the Southern Hemisphere makes up more of the recorded events.

From teh same column we can look at the Severity Plots for the Total Time Period. The Middle Severity of 1.5 does not start until 2005. This suggests that the rating for this category did not previously exist, judging by how much it is currently used.

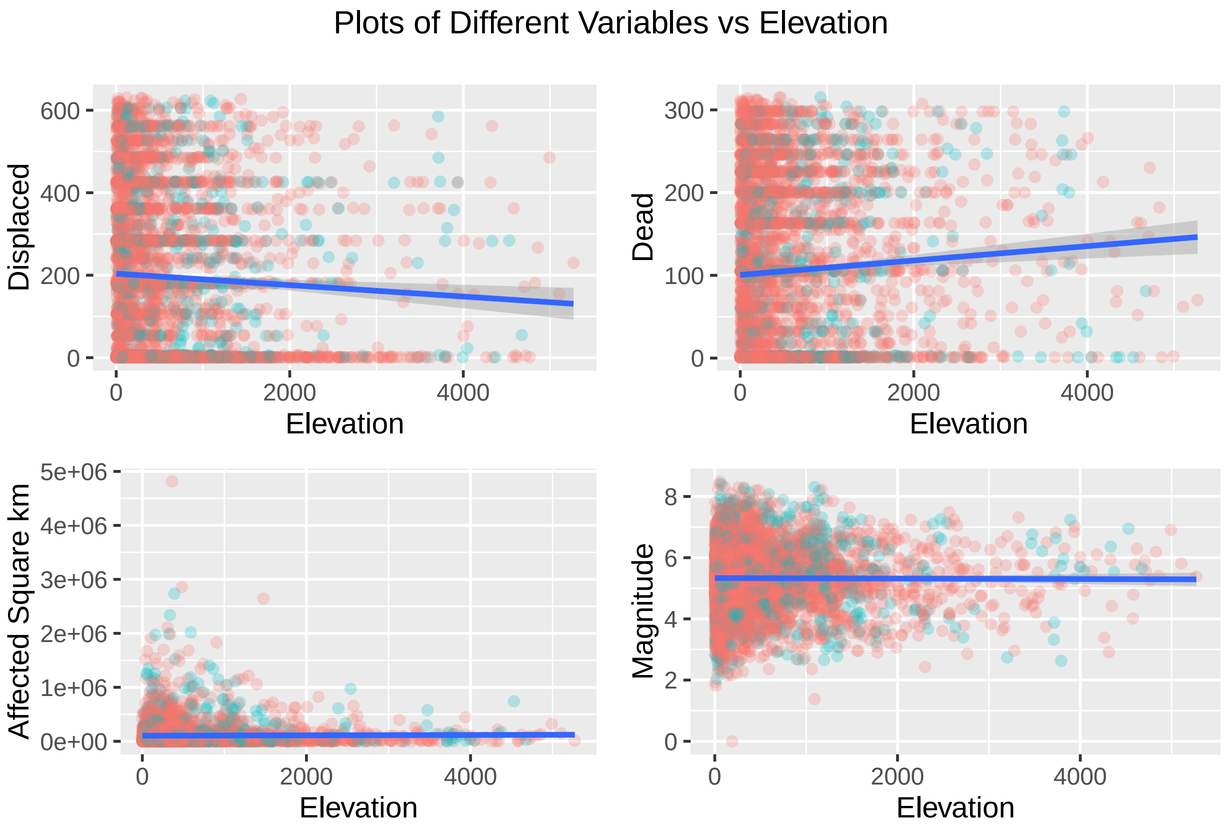
We then began to look for correlations between elevation and different aspects of the flooding events. The general intution would be that the higher the elevation, the lower the number of flood events

As we look at the right side of the plots "Groupd by Month" we can further see some interesting characteristics. Looking at the upper right plot it would seem that for the world most of the flooding occurs in the 7th month, July. However as we break this down to Northern and Southern Hemispheres we can see that these seasons are opposite for each region. This makes sense because the different hemispheres experience opposite seasons.

Severity per month does not tell us much more additional information, the splits are very similar.



An interesting finding fomr this graph is the outlier that was located at the top for Duration in Days. We can see that there is a flood event that happend in the USA and lasted for 419 days. The next closest was 168 days which occured in Zambia. We can also see a negative correlation with the elevation and duration in days. This was removed in the plotting due to resolution.



From these plots we can notice that there is a negative correlation between elevation and the number of displaced people for the flood events. In contract see a positive correlation between elevation and the number of dead.

For Affected Square km it might seem that there is no correlation but there is a slightly negative correlation between teh Elevation and Affected Square km. the Y-axis is suppressing this but is not a strong enough correlation to be relevant. #

2003 Flood Event Investigation

First we will take a holistic view of the world map during 2003, which was the year that generated the largest number of floods. While this could have been just due to reporting we will focus on this year to continue to narrow our scope of the data.

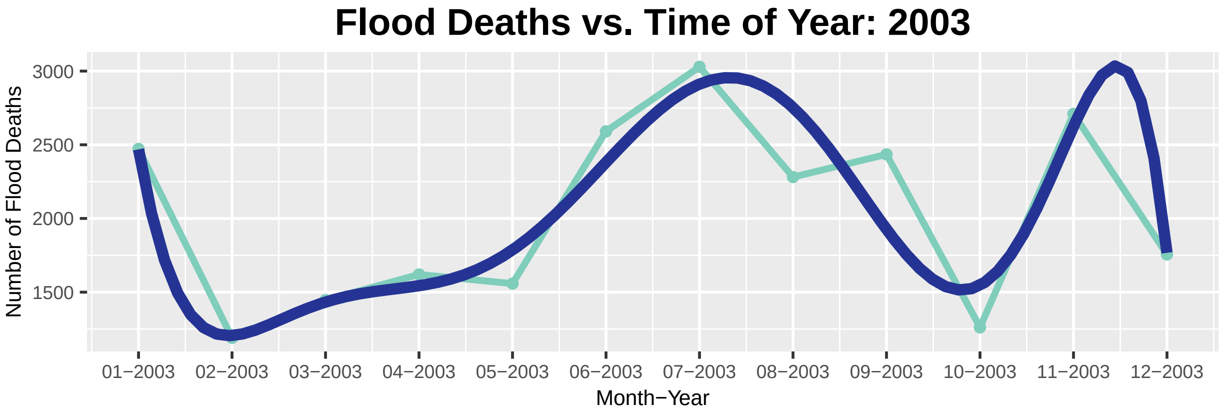
Displayed below is an interactive world map showing the number of deaths by country in 2003. The countries with fewer deaths are lighter and the countries with more deaths are darker. Run the mouse over a country to display its name and number of deaths associated with it.

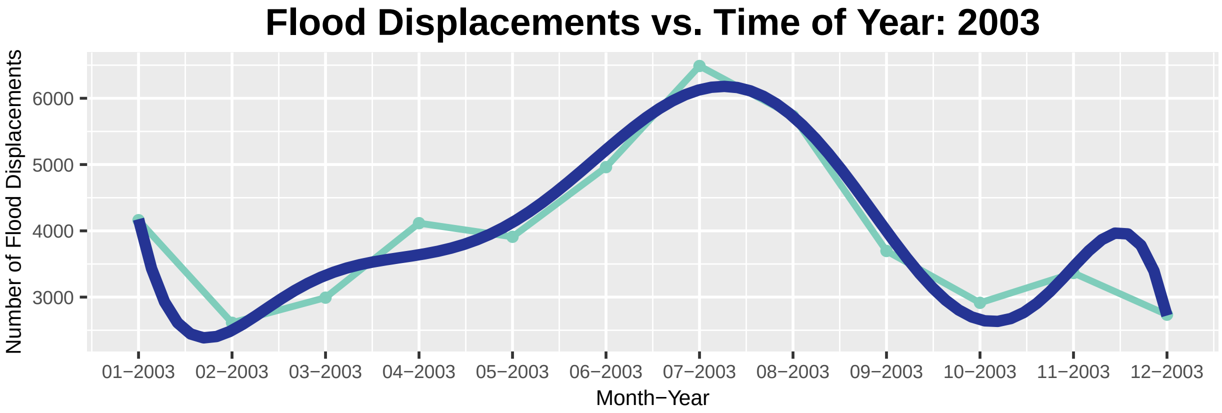
World Map - Deaths in 2003

Next is another interactive World map showing the number of displaced individuals in 2003. It has the same format and style as the previous one; run the mouse over the countries to see the country name and magnitude.

World Map - Displacements in 2003

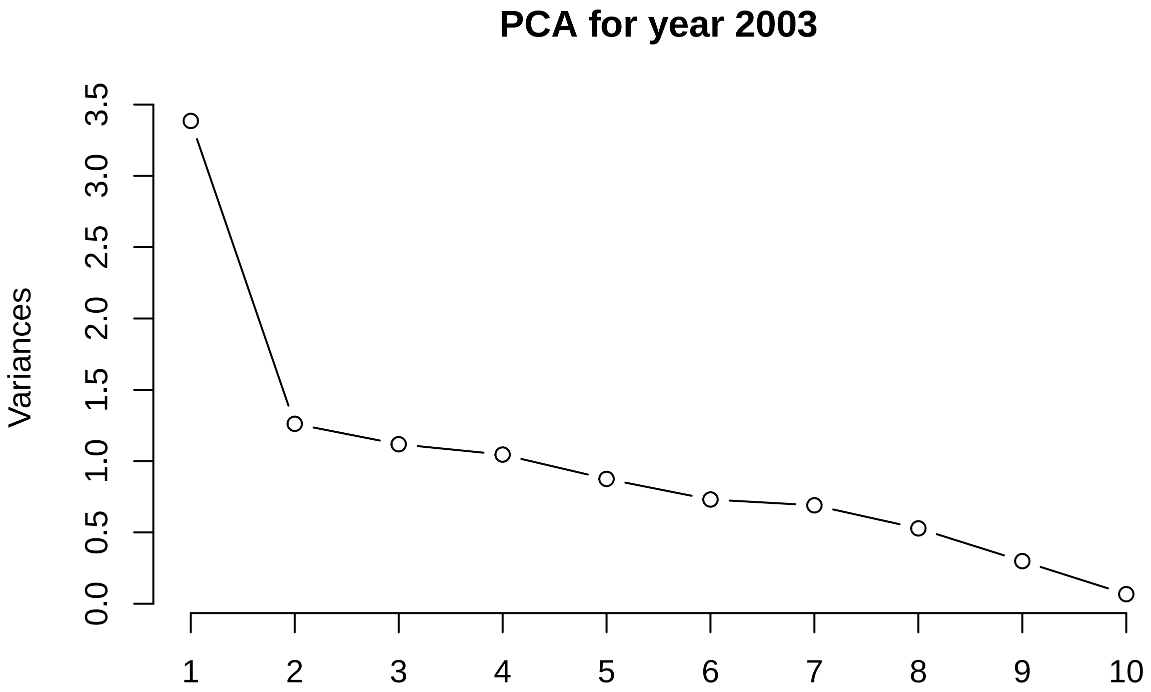
By looking at 2003 more closely, we can plot the number of deaths and displacements by month in that year. Both plots showed a similar overall trend; the number of deaths and displacements both reached their maximums from June to August. The faded turquoise line maps the actual number of deaths and displacements, while the dark blue smooth curve is an attempt to represent the overall trend in a more visually appealing way.





We did PCA analysis on the 2003 data, with 10 numerical variables of “Duration.in.Days”, “Dead”, “Displaced”, “Severity..”, “Affected.sq.km”, “Magnitude..M. . . ”, “Centroid.X”, “Centroid.Y”, “M.6” and “M.4”. And we grouped by the Main Causes of the floods.

In 2003, there are 7 main causes, which are “Heavy rain”, “Tropical Cyclone”, “Brief torrential rain”, “Dam/Levy, break or release”, “Monsoonal rain”, “Snow Melt”, “Ice Jams”, “Rain and Snow Melt” and “Rainy Season”.



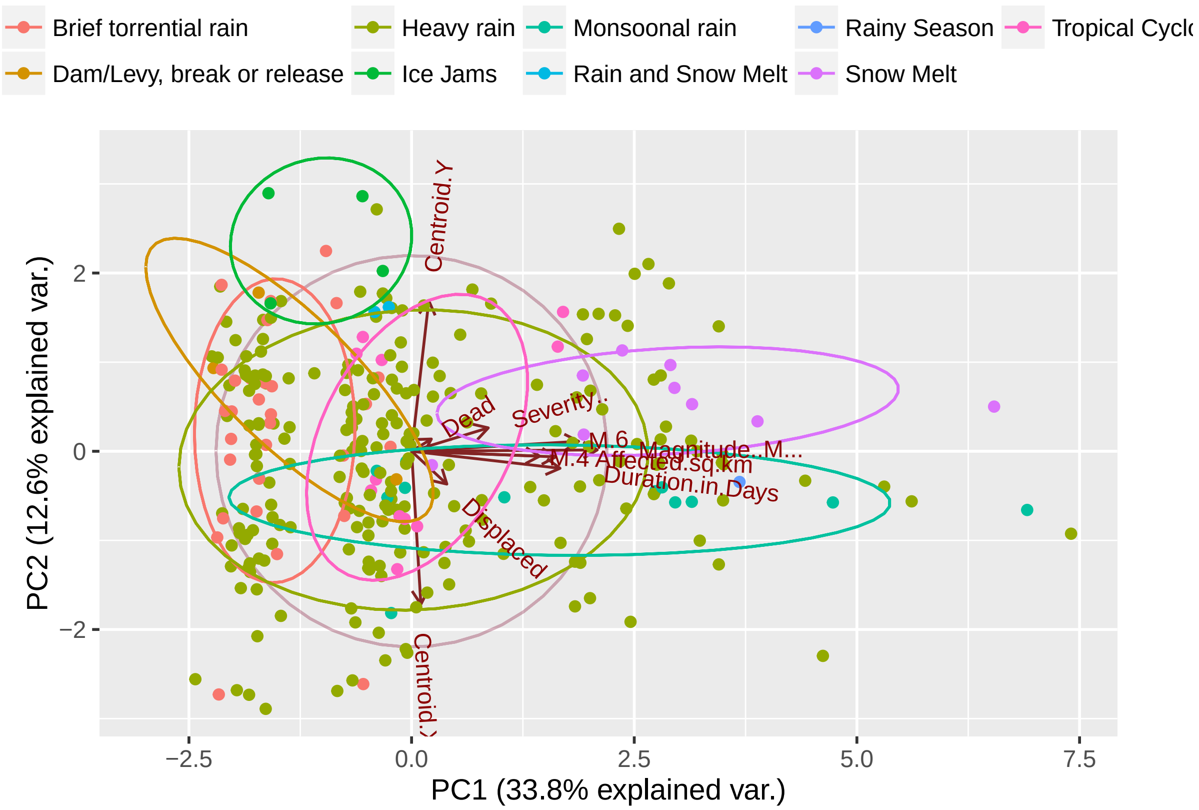
There is one Priciple Component that explains quite a lot of the variance. The left PCs drop quickly after first one and decrease accordingly.

Let's look at the summary of the PCA.

```
## Importance of components:
##                           PC1    PC2    PC3    PC4    PC5     PC6    PC7
## Standard deviation     1.8398 1.1231 1.0574 1.0226 0.9354 0.85483 0.8307
## Proportion of Variance 0.3385 0.1261 0.1118 0.1046 0.0875 0.07307 0.0690
## Cumulative Proportion  0.3385 0.4646 0.5765 0.6810 0.7685 0.84161 0.9106
##                            PC8     PC9    PC10
## Standard deviation     0.72677 0.54655 0.25895
## Proportion of Variance 0.05282 0.02987 0.00671
## Cumulative Proportion  0.96342 0.99329 1.00000
```

The first PC consists of 33.85% of the total variance, together with the second PC, 46% of the variance can be explained.

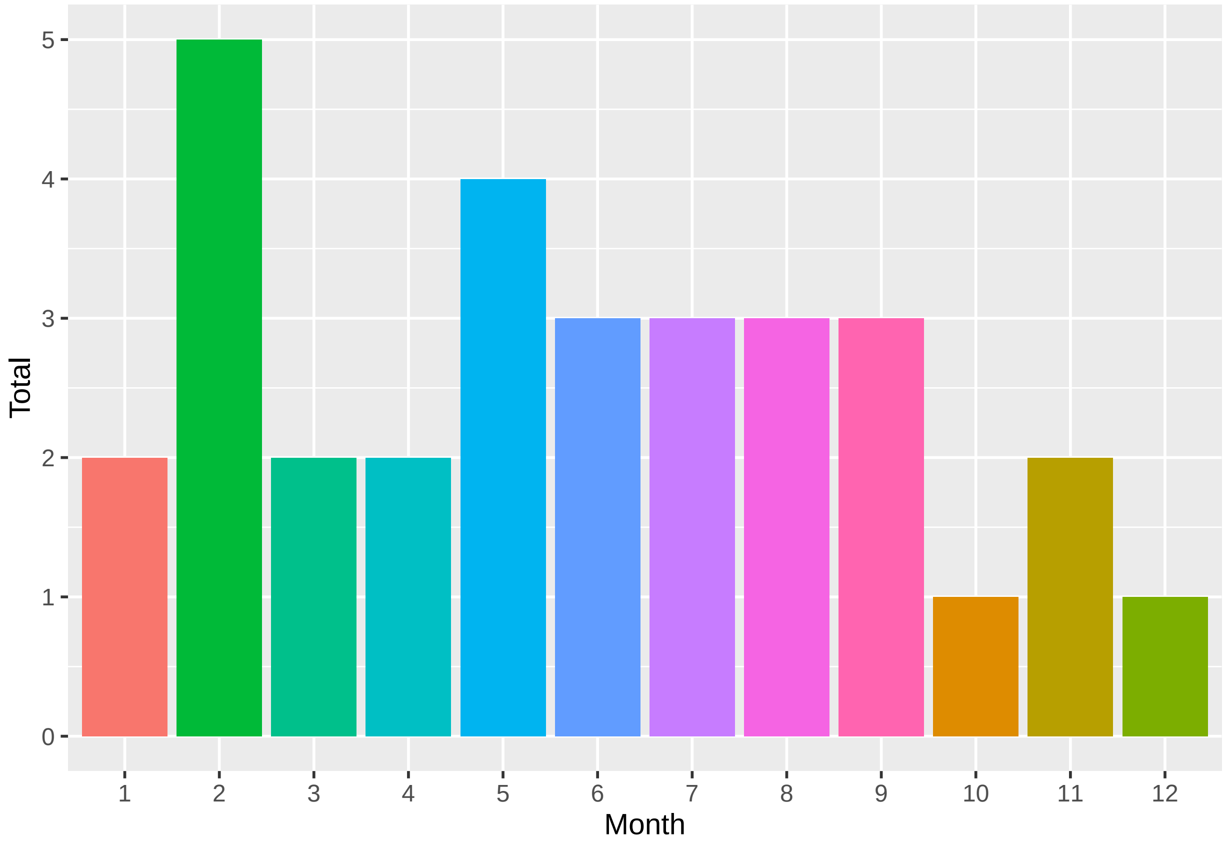
Next let's look at the groups by first two principle components.



The plot shows except two centroid, the other 8 factors have similar directions and they can be explianed by first principle component well. Among the groups of the main causes of floods, we can see "Snow Melt" and "Monsoonal rain" typically lie more right, which indicates heavier damage. While "Ice Jams" and "Brief torrential rain" are left in the plot, which indicates less damage.

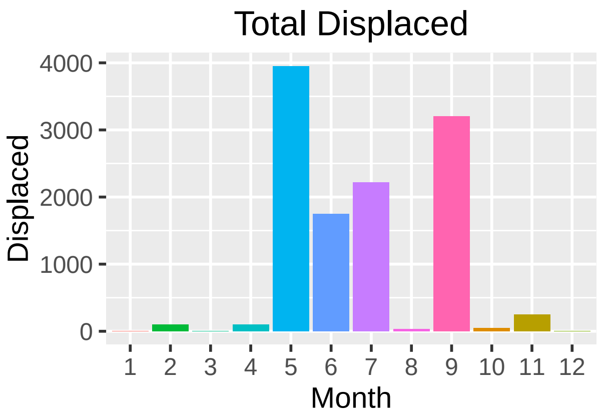
2003 USA Flood Analysis

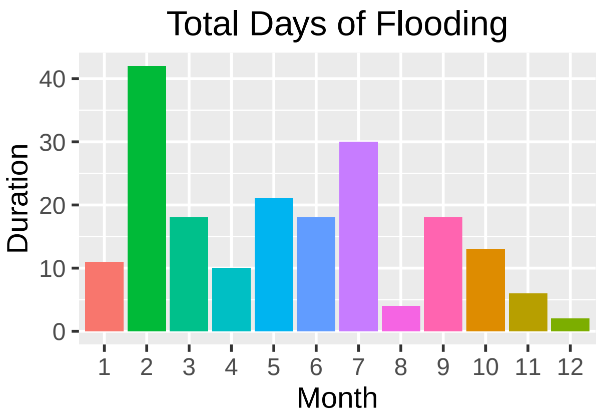
After some general analysis, we noticed that the most floods took place in 2003. Of all the events that occurred that year, the USA had recorded the most floods. Below is a plot of floods by month in the USA in 2003.

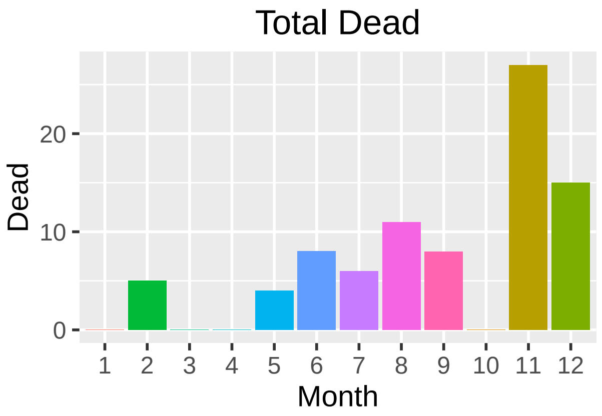


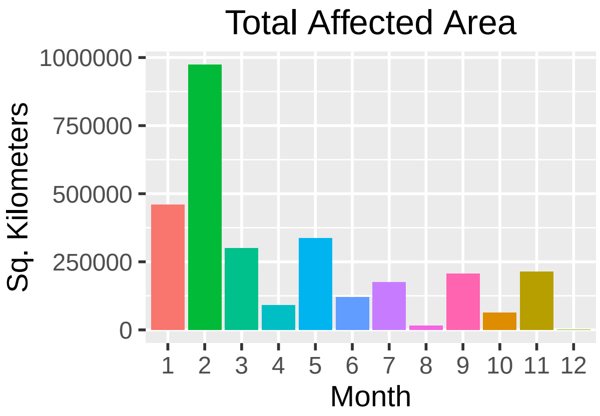
Clearly, the most floods occured in February and there was at least one flooding event in every month. Next, we examine the data describing the damage caused by floods in the USA.

Damage by Month in USA, 2003









As mentioned above, February was the month that had the most recorded floods (5). As would be expected, February also had the most flooding days (around 40) and largest cumulatively affected area by a significant margin. Considering these facts, it is surprising that the total number of people displaced was very low, especially compared to the total number of people displaced in May, June, July, and September. Interestingly, the total number of dead is relatively low compared to other months that had fewer floods in a smaller total area.

One might assume that this is because the magnitude of these floods was less, but in fact all but one of the floods occurring in February ranked above 5.064516 (average flood magnitude in 2003). Another explanation for this is where the floods occurred. The February floods were in southern California, eastern Kentucky, the southern Mid-Atlantic, along the Gulf Coast (Mississippi and Louisiana), and in Virginia. The only area where nobody was displaced was Eastern Kentucky, which is also where two people were recorded dead. One possible explanation is that these locations were less populated than some of the other flood locations. It is also possible that these locations were more prepared for the floods and therefore incurred fewer problems and casualties. Unfortunately, there was no data recorded for the damage in US dollars for these floods.

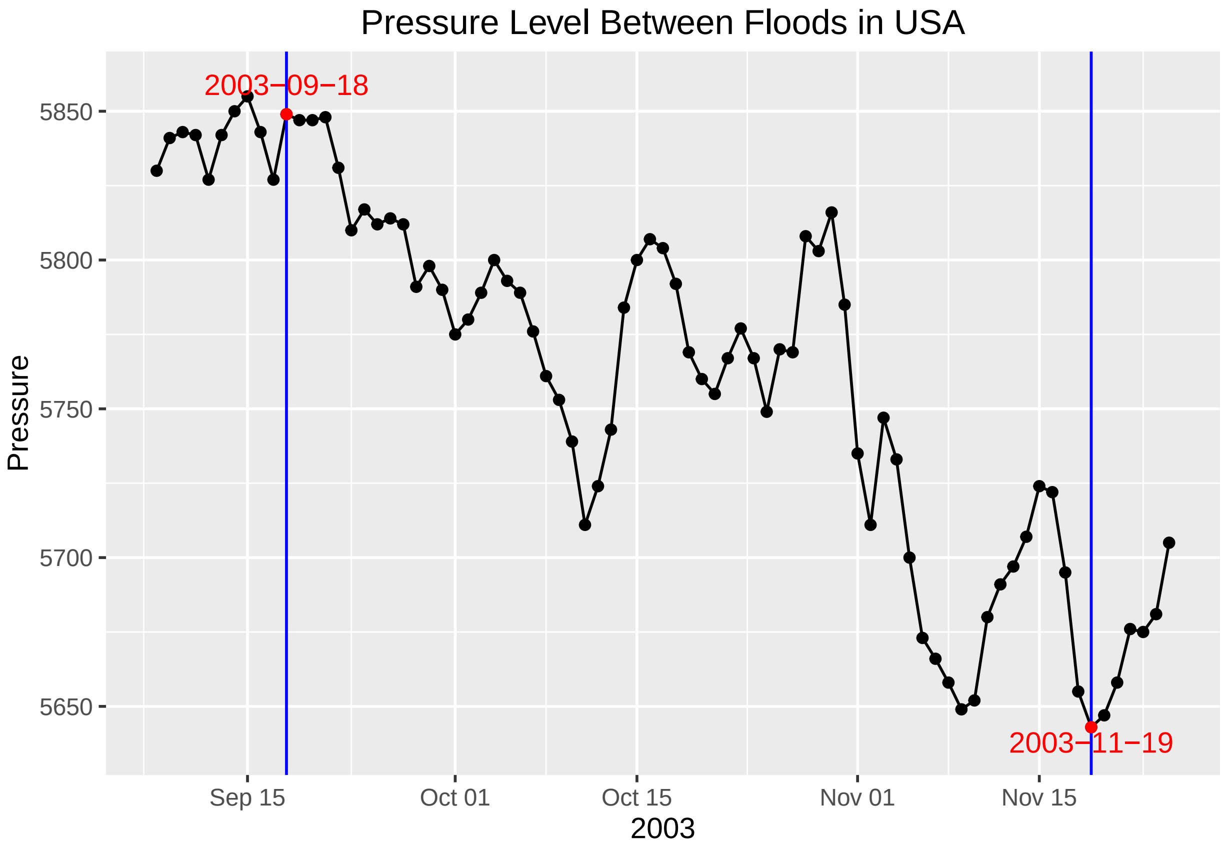
Pressure Levels Between Floods

We decided it might be interesting to find a location where several events occured withing the span of a few months and reference it with the day-by-day pressure to see if there was a pattern. We looked at floods that began on September 18, 2003 and November 19, 2003 (ending on September 24, 2003 and November 22, 2003 respectively). The main cause of the September flood was a tropical cyclone and heavy rain was the main cause of the November event. The magnitudes and locations of these floods was 6.1, 6.2 and (-78.4245, 37.3311), (-79.3157, 38.8759) respectively. These coordinates correspond to a point in the mid-Atlantic; more specifically, parts of Maryland, West Virginia, and Virginia were the areas affected by these events.

In the following plots the events in red circular markers signify the flood event in the US corresponding with the annotated events above.

2003-09-18 Flood Event in red, Duration: 7 Days   2003-11-19 Flood Event in red, Duration: 4 Days

The relationship we can see between these two plots for the event occuring in the USA,indicated by the red data point, is the flooding seems to begin when a low pressure system comes into contact with a higher pressure system and this starts the flooding event.



One noticeable feature of the graph is that the pressure level seemed to have trended downward from September to November. One possible explanation for this could be seasonal variation. By examining pressure levels in the days leading up to the beginning of the flood, one can observe a sharp change pressure levels. Starting on September 15th, the pressure levels drops sharply and then increases sharply. Starting on November 16th, the pressure levels also drop sharply in the days leading up to the beginning of the flood.

At first glance, these findings might seem interesting, but when one examines the pressure levels in the days between the two floods, there are many sharp increases and decreases. Since there were no other recorded floods in the region during this time frame, one must proceed with caution when attributing the cause of these floods to changes in pressure levels leading up to the events. If anything, the pressure level was a contributing factor to the heavy rains in these regions.

These visuals allows us to think about other possible approaches to what determines a flooding event.The data point is a single coordinate that came with the data but we do have the ability to expand that into a flood region. This might give more insight into what was affected.

There is also other factors to consider such as rivers, water shelfts and lakes around the area. It could be that there was heavy rain in other areas that casued downstream flooding of rivers. These visuals do seem to point to flooding as a result of change in pressure but more investigation is required.

Conclusion

The exploration of the global flood data set was conducted in a hierarchical manner; starting with broader analysis, we slowly shifted our focus to a specific year and eventually singular events. The initial analysis and visualization was performed on all floods ranging from 1985-2015 by using several tools such as interactive world maps (where events are pinned), word clouds, bar charts and density plots. After the preliminary analysis, we decided to delve deeper into the year with the most flood events: 2003. World maps for the number of deaths and displacements (in 2003) were created and Principal Components Analysis was performed to get a better understanding of the relationship between variables. Next, we decided to further refine our analysis by only including flood events in the USA. A more general interactive feature was also added: a search function which displays the floods by year (or range of years) and area affected. Our general methodology was to start broad and eventually focus on several subsets of the data but it is important to mention that despite conducting a comprehensive analysis, an exhaustive analysis on a data set this size was not within the scope of the project.