Reproducible Research: Storm Data Analysis

May 1, 2017

NOAA Storm Data: Health and Economic Effects

Introduction

This project analyses information collected by the National Weather Service. The data is first cleaned and tidyed, and then used to determine which types of storms/weather events are most hazardous in terms of population health or damage costs. Finally, the distribution of certain storm types accross the US are shown, indicating which areas are most at risk.

```
opts_chunk$set(fig.width=8, fig.height=8,dpi=144)
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(ggmap)
```

```
## Loading required package: ggmap
```

obtaining data

The data for this analysis comes from the National Oceanic and Atmospheric Administration's National Weather Service. The (compressed) csv can be downloaded from here. The data for this analysis was downloaded on February 28, 2017.

```
destfile='StormData.csv.bz2'
if (! file.exists(destfile))
{
    download.file('https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2'
}
stormData<-read.csv('./StormData.csv.bz2',stringsAsFactors=FALSE)
```

```
str(stormData)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951
##  $ BGN_TIME  : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE : chr  "CST" "CST" "CST" "CST" ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI   : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ LENGTH    : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH     : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F         : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ LATITUDE  : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS   : chr  "" "" "" "" ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

Data Processing

The Event type code (EVTYPE) is important, as it will be used to categorise the data. Unfortunately, there are many errors and inconsistencies in the raw data.

```
length(unique(stormData$EVTYPE))
```

```
## [1] 985
```

```
moose<-unique(tolower(stormData$EVTYPE))
length(moose)
```

```
## [1] 898
```

```
moose[grep('aval',moose)]
```

```
## [1] "avalanche"                     "avalance"
## [3] "heavy snow/blizzard/avalanche"
```

```
sum(grepl('tstm',moose))
```

```
## [1] 29
```

```
sum(grepl('thunderstorm',moose))
```

```
## [1] 81
```

There are 87 event types that differ only in the case (upper or lower) of their lettering, ‘avalanche’ is missspelled at least once, and high winds may be categorised as something like ‘tstm wind’ or ‘thunderstorm wind’, with variations based on wind speeds/gust speeds. All of these variations make it difficult to compare weather events. According to the instructions on storm data preparation, there are only 48 allowed event types, which are listed below:

- Astronomical Low Tide
- Avalanche
- Blizzard
- Coastal Flood
- Cold/Wind Chill
- Debris Flow
- Dense Fog
- Dense Smoke
- Drought
- Dust Devil
- Dust Storm
- Excessive Heat
- Extreme Cold/Wind Chill
- Flash Flood
- Flood
- Frost/Freeze
- Funnel Cloud
- Freezing Fog
- Hail
- Heat
- Heavy Rain
- Heavy Snow
- High Surf
- High Wind
- Hurricane (Typhoon)
- Ice Storm
- Lake-Effect Snow
- Lakeshore Flood
- Lightning
- Marine Hail
- Marine High Wind
- Marine Strong Wind
- Marine Thunderstorm Wind
- Rip Current
- Seiche
- Sleet
- Storm Surge/Tide
- Strong Wind
- Thunderstorm Wind
- Tornado
- Tropical Depression
- Tropical Storm

- Tsunami
- Volcanic Ash
- Waterspout
- Wildfire
- Winter Storm
- Winter Weather

The raw data will need to be tidied before it can be processed effectively. First the observations (rows) of interest are selected (as this will limit the range of EVTYPE codes that need to be corrected), and then regular expressions are used to correct the event type codes to one of the 48 options listed above. Events that don't correspond to these event types will be removed, as they are not "allowed" by the NOAA and so may not have been recorded consistently (the events of these types that are included may not give an accurate representation of these storm types).

Based on this post in the course discussion forums, it was only after January 1996 that NOAA started recording events of all types. Tornado data was present from the beginning (1950), but other types of weather events were reported and recorded later. Data prior to January 1996 will be omitted, as analysis of this data could introduce bias due to lack of records on certain weather types.

```
library(dplyr)
stormData$BGN_DATE<-as.Date(stormData$BGN_DATE,format='%m/%d/%Y')
storm96<-stormData %>% filter(BGN_DATE >  '1996-01-01')
```

Crop and property damage are stored strangely, with the first few significant digits stored seperately from the dollar exponent. The actual cost are calculated by multiplying the values in the damage column by one thousand, one million, or one billion for exponent values of "K","M", or "B" respectively.

As this the object of the analysis is to study the economic and health effects of weather events, observations in which there were no injuries or fatalities, or damage to crops or property, are of little interest. These observations are removed.

```
storm96$CropDamage<-storm96$CROPDMG
levels(as.factor(storm96$CROPDMGEXP) )
```

```
## [1] ""  "B" "K" "M"
```

```
# are the '' values in the CROPDMGEXP relevant?
min(storm96$CROPDMG[storm96$CROPDMGEXP==''])
```

```
## [1] 0
```

```
max(storm96$CROPDMG[storm96$CROPDMGEXP==''])
```

```
## [1] 0
```

```
# ...no

# scale crop damage
thelist<-with(storm96, CROPDMGEXP=='B')
storm96$CropDamage[thelist]<-storm96$CropDamage[thelist]*1.0e9
thelist<-with(storm96, CROPDMGEXP=='K' )
storm96$CropDamage[thelist]<-storm96$CropDamage[thelist]*1.0e3
thelist<-with(storm96, CROPDMGEXP=='M')
storm96$CropDamage[thelist]<-storm96$CropDamage[thelist]*1.0e6

# scale property damage
storm96$PropDamage<-storm96$PROPDMG
levels(as.factor(storm96$PROPDMGEXP) )
```

```
## [1] ""  "0" "B" "K" "M"
```

```
# are the '0' values in the PROPDMGEXP relevant?
min(storm96$PROPDMG[storm96$PROPDMGEXP=='0'])
```

```
## [1] 0
```

```
max(storm96$PROPDMG[storm96$PROPDMGEXP=='0'])
```

```
## [1] 0
```

```
# ...no

thelist<-grepl('[bB]',storm96$PROPDMGEXP)
storm96$PropDamage[thelist]<-storm96$PropDamage[thelist]*1.0e9
thelist<-grepl('[mM]',storm96$PROPDMGEXP)
storm96$PropDamage[thelist]<-storm96$PropDamage[thelist]*1.0e6
thelist<-grepl('[kK]',storm96$PROPDMGEXP)
storm96$PropDamage[thelist]<-storm96$PropDamage[thelist]*1.0e3

# slim data based on injuries, fatalities, and property/crop damage
slim96<- storm96 %>% select(BGN_DATE,EVTYPE,PropDamage,CropDamage,INJURIES,FATALITIES,LATITU
filter(PropDamage>0 | CropDamage >0 | INJURIES >0 | FATALITIES >0)
length(unique(slim96$EVTYPE))
```

```
## [1] 222
```

Right now there are 222 event types. The final dataset should only contain some subset of the allowed event types listed above. Start by casting everything to lower case, and by removing any leading whitespace.

```r
slim96$EventType<-tolower(slim96$EVTYPE)
# trimws is new? doesn't exist in older R versions
#slim96$EventType<-trimws(slim96$EventType)
slim96$EventType<-gsub('^\\s+|\\s+$','',slim96$EventType)
length(unique(slim96$EventType))
```

```
## [1] 183
```

Regular expressions (implemented through the grepl and gsub functions) are used to rename event types.

```r
# remove things that can't be reclassified easily
badcodes<-c("astronomical high tide","other","marine accident", "coastal storm","coastalstor
slim96<- slim96 %>% filter(! (EventType  %in% badcodes))

# change non thunderstorm winds to strong winds
slim96$EventType<-gsub('non[ -]tstm wind','strong wind',slim96$EventType)

#change tstm/thunderstorm wind to thunderstorm wind
change<-with(slim96, (grepl('tstm',EventType) | grepl('thunderstorm',EventType)) & grepl('wi
slim96$EventType[change]<-'thunderstorm wind'

# change 'blowing dust' to dust devil
slim96[slim96$EventType=='blowing dust','EventType']<-'dust devil'

#change mudslide or similar to debris flow
slim96$EventType<-gsub('mud[ -]?slides?','debris flow',slim96$EventType)

# change any remaining occurences of "coastal" to coastal flood
slim96[grepl('coastal|tidal|cstl',slim96$EventType),'EventType'] <-'coastal flood'

# change (non extreme) cold to cold/windchill
slim96[grepl('cold|thermia',slim96$EventType) & (! grepl('extreme',slim96$EventType)),'Event

# extreme cold
slim96$EventType[grepl('extreme',slim96$EventType)] <-'extreme cold/wind chill'

# land or rock slides to debris flow
slim96$EventType[grepl('landsl|rock',slim96$EventType)] <-'debris flow'

# fog to dense fog
```

```
slim96[grepl('fog',slim96$EventType),'EventType'] <-'dense fog'

# blowing dust or whirlwind to dust devil
slim96$EventType[grepl('whirl|dust',slim96$EventType)]<-'dust devil'

# record heat to excessive heat
slim96$EventType[grepl('record heat',slim96$EventType)]<-"excessive heat"

# flash floods
slim96$EventType[grepl('flash',slim96$EventType)]<-"flash flood"

# non flash floods
slim96$EventType[grepl('high water|flood|fld',slim96$EventType) & (! grepl('flash|coastal',s

# frost/freeze
slim96$EventType[grepl('agric|black|road|free?z|frost',slim96$EventType) ]<-"frost/freeze"

# "landspout"
slim96$EventType[grepl('landspout',slim96$EventType)]<-"funnel cloud"

# hail
slim96$EventType[grepl('hail',slim96$EventType)]<-"hail"

# excessive heat
slim96$EventType[grepl('record',slim96$EventType)]<-"excessive heat"

# regular heat
slim96$EventType[grepl('heat|warm',slim96$EventType) & (! grepl('excessive',slim96$EventType

# rain (not gusty or freezing)
slim96$EventType[grepl('rain',slim96$EventType) & (! grepl('wind|freez',slim96$EventType))]<

# lake effect snow
slim96$EventType[grepl('lake effect snow',slim96$EventType)]<-"lake-effect snow"

# winter weather
slim96$EventType[grepl('blowing|falling|light|mix',slim96$EventType)]<-"winter weather"

# heavy snow
slim96$EventType[grepl('snow',slim96$EventType) & (! grepl('lake',slim96$EventType))]<-"heav

# high surf
slim96$EventType[grepl('surf|sea|wave|swell',slim96$EventType)]<-"high surf"

# high wind
slim96$EventType[grepl('grad|high wind',slim96$EventType)]<-"high wind"
```

```r
# strong wind
slim96$EventType[grepl('gust|wind damage',slim96$EventType)]<-"strong wind"
slim96$EventType[ slim96$EventType =='wind' ]<-'strong wind'
slim96$EventType[ slim96$EventType=='winds']<-'strong wind'
slim96$EventType[ slim96$EventType== 'strong winds']<-'strong wind'

# hurricane (typhoon)
slim96$EventType[grepl('hurricane|typhoon',slim96$EventType)]<-"hurricane (typhoon)"

# rip currents (assume that drowning belongs here)
slim96$EventType[grepl('rip current|drown',slim96$EventType)]<-"rip current"

# storm surge
slim96$EventType[grepl('storm surge',slim96$EventType)]<-"storm surge/tide"

# thunderstorm winds - microbursts/downbursts are often associated with thunderstorms, so th
slim96$EventType[grepl('burst',slim96$EventType)]<-'thunderstorm wind'
slim96$EventType[ slim96$EventType== 'thunderstorm']<-'thunderstorm wind'

# wildfire
slim96$EventType[grepl('fire',slim96$EventType)]<-'wildfire'

#print the current list of event types
moose<-unique(slim96$EventType)
moose[order(moose)]
```

```
##  [1] "astronomical low tide"   "avalanche"
##  [3] "blizzard"                "coastal flood"
##  [5] "cold/wind chill"         "debris flow"
##  [7] "dense fog"               "dense smoke"
##  [9] "drought"                 "dust devil"
## [11] "excessive heat"          "extreme cold/wind chill"
## [13] "flash flood"             "flood"
## [15] "frost/freeze"            "funnel cloud"
## [17] "hail"                    "heat"
## [19] "heavy rain"              "heavy snow"
## [21] "high surf"               "high wind"
## [23] "hurricane (typhoon)"     "ice storm"
## [25] "lake-effect snow"        "marine strong wind"
## [27] "rip current"             "seiche"
## [29] "storm surge/tide"        "strong wind"
## [31] "thunderstorm wind"       "tornado"
## [33] "tropical depression"     "tropical storm"
## [35] "tsunami"                 "volcanic ash"
## [37] "waterspout"              "wildfire"
## [39] "winter storm"            "winter weather"
```

The following event codes are absent from the data: 1. dust storm 2. freezing fog 3. lakeshore flood 4. lightning 5. marine hail 6. marine high wind 7. marine thunderstorm wind 8. sleet

The final, tidied data used for this analysis is stored in the dataframe "slim96".

Results

Do the figure(s) have descriptive captions (i.e. there is a description near the figure of what is happening in the figure)? Does the analysis address the question of which types of events are most harmful to population health? Does the analysis address the question of which types of events have the greatest economic consequences? ** axis labels and units on all plots **

The event types of interest are those which have the largest effects on population health or which have the highest costs in terms of property or crop damage. Instead of looking at individual storm events, it is useful to summarise the data based on the weather event type, as this gives a better idea of which storm types were most significant overall.

```
#Form total damage from crop and property damage
slim96$TotalDamage<-slim96$CropDamage + slim96$PropDamage

#aggregate data based on EventType (total)
AggStorms<- slim96 %>% select(EventType,TotalDamage,FATALITIES,INJURIES) %>% group_by(EventT
summarise_each(funs(sum),TotalDamageSum=TotalDamage,TotalInjuries=INJURIES,TotalFatalities=F
```

Across the United States, which types of events are most harmful with respect to population health?

From the aggregated data, the number of casualties is computed for each event type as the sum of the total number of injuries and fatalities. The casualties for the 20 most hazardous event types are plotted below.

```
# get the 20 storm types with the highest total casualties (fatalities + injuries)
casStorms<-AggStorms %>% mutate(casualties=TotalInjuries + TotalFatalities) %>% arrange(desc

# plot this in a bar chart
g<-ggplot(data=casStorms,aes(x=EventType,y=casualties,fill=EventType))
g+ geom_bar(stat='identity') + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + gu
```

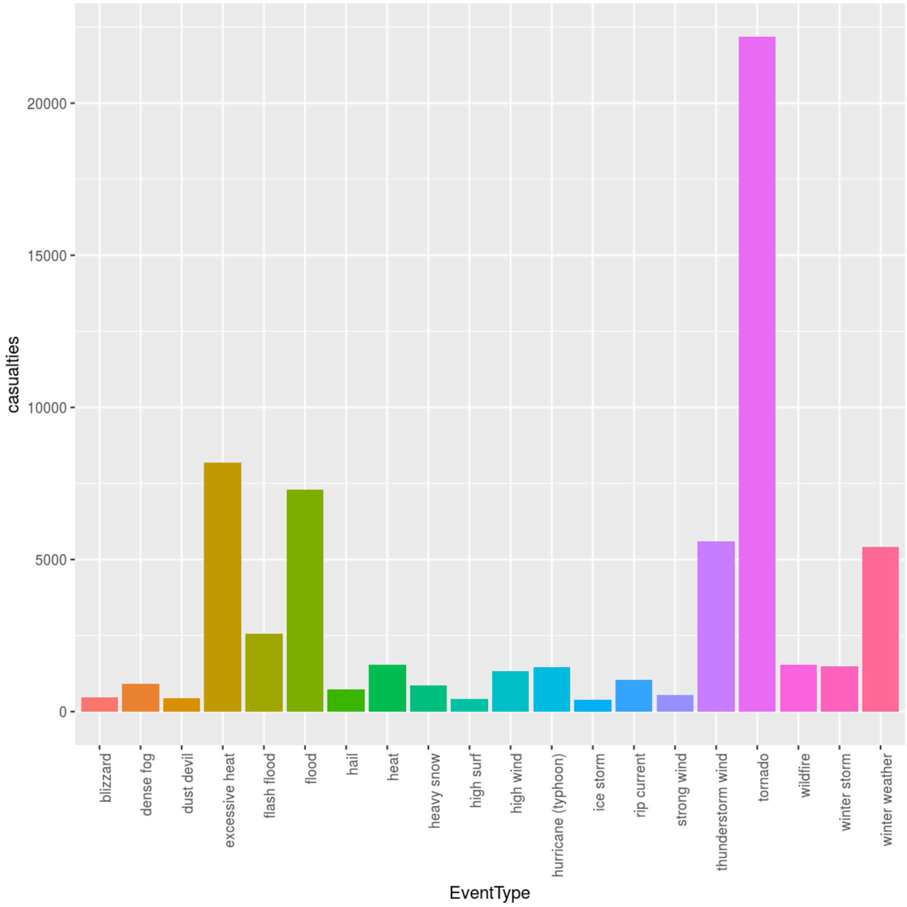

Figure 1: plot of chunk plot1

Tornadoes are the most dangerous type of storm in terms of population health, with more than twice as many casualties as either floods or excessive heat. ### Across the United States, which types of events have the greatest economic consequences?

What are the costliest (individual) events? Print the 10 storms with the highest damage (property + crop) costs.

```
slim96 %>% select(BGN_DATE,EventType,TotalDamage,LATITUDE,LONGITUDE) %>% arrange(desc(TotalD
```

```
##      BGN_DATE           EventType  TotalDamage LATITUDE LONGITUDE
## 1  2006-01-01               flood 115032500000     3828     12218
## 2  2005-08-29    storm surge/tide  31300000000        0         0
## 3  2005-08-28 hurricane (typhoon)  16930000000        0         0
## 4  2005-08-29    storm surge/tide  11260000000        0         0
## 5  2005-10-24 hurricane (typhoon)  10000000000        0         0
## 6  2005-08-29 hurricane (typhoon)   7390000000        0         0
## 7  2005-08-28 hurricane (typhoon)   7350000000        0         0
## 8  2004-08-13 hurricane (typhoon)   5705000000        0         0
## 9  2001-06-05      tropical storm   5150000000        0         0
## 10 2004-09-04 hurricane (typhoon)   4923200000        0         0
```

Many of these events are from the 2005 Atlantic hurricane season Hurricane Katrina was active during August 23-29, 2005, which accounts for 5 of 10 storms on this list. The October 24 storm (#5) corresponds to hurricane Rita, and the the 150 billion flood dated January first, 2006 is also probably Katrina related.

Below is a bar chart of the 20 most damaging storm types.

```
damageStorms<-AggStorms %>% arrange(desc(TotalDamageSum)) %>% head(n=20)

g<-ggplot(data=damageStorms,aes(x=EventType,y=TotalDamageSum,fill=EventType))
g+ geom_bar(stat='identity') + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + gu
labs(title='Damage Costs for Weather Events',y='Damage ($)',x='Event Type')+
labs(title='Casualties for Weather Events',y='Casualties',x='Event Type')
```

Floods are the most damaging event type, with a total cost of almost $150 billion USD. Note that most of this ($115 billion) is due to the flood on Jan 1, 2006. As mentioned earlier, this is presumably an effect of hurricane Katrina.

Which areas are most prone to floods and tornadoes?

(this wasn't an assigned question, but I thought it would be an interesting plot)

The storm dataset contains information on lattitude and longitude. A map of the US can be displayed using the ggmap package, which works in conjunction with

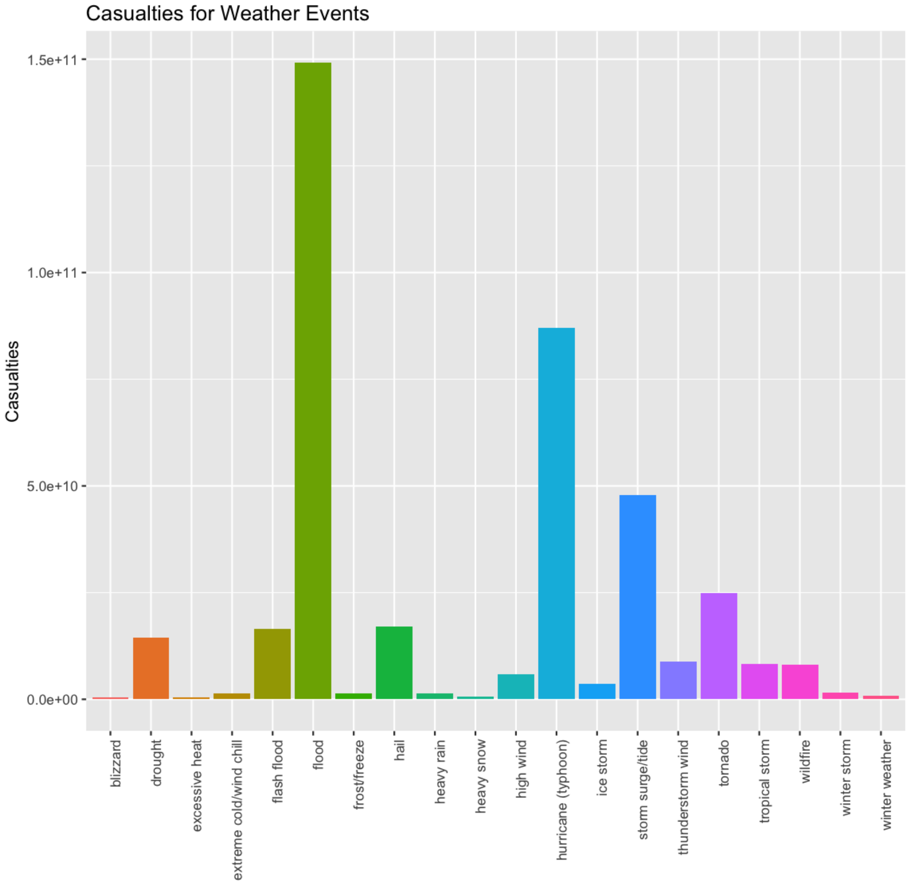

Figure 2: plot of chunk plot1a

the ggplot2 plotting framework. The location data can then be overlaid on top of the map. A scatter plot could be used, but there are a lot of overlapping data points that make interpreting the data difficult. Instead, the location information is used to define a density function, and the contours of this function are plotted.

(plots are made using ggmap and ggplot2)

```
#
floodnados<-slim96 %>% filter(EventType %in% c('flood','tornado')) %>% mutate(Lat=LATITUDE/1
# not 100% sure that the lat/long mapping is correct. Think that the Longitude needs that fa
mapus2<-ggmap(get_map("usa",zoom=4),extent='device')
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=usa&zoom=4&size=640x6
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=usa&senso
```

```
## Warning: `panel.margin` is deprecated. Please use `panel.spacing` property
## instead
```

```
mapus2 + stat_density2d(data=floodnados,aes(x=Lon,y=Lat,fill=..level..,alpha=..level..,colou
labs(title='Flood and Tornado distribution across the US')
```

```
## Warning: Removed 4517 rows containing non-finite values (stat_density2d).
```

Both types of weather events are on the Eastern half of the US. Flods seem to be more localised, generally occuring beneath the great lakes, but with a very high density in Iowa. Tornadoes are spread more evenly over the south eastern US, with the highest concentration in south (Alabama, Mississippi, Louisiana, Georgia).

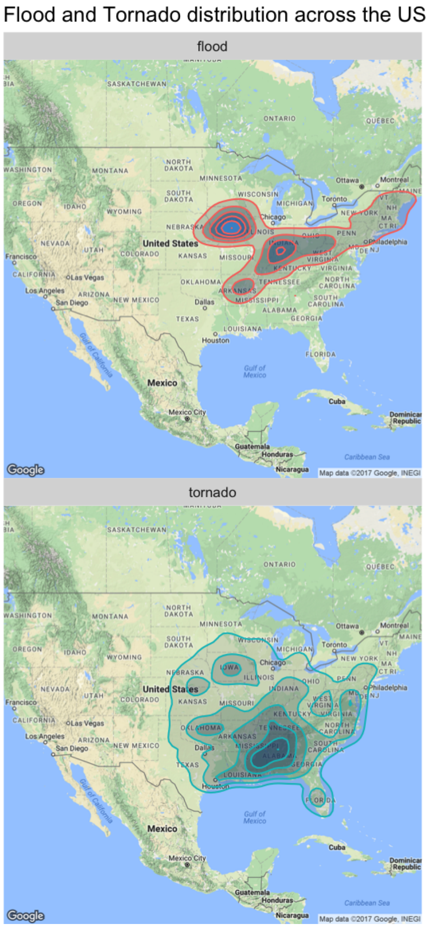

Figure 3: plot of chunk mapplotsa