Coursera Reproducible Research: Course Project 2

Chamika Senanayake

9/14/2020

Exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database

Synopsis

This project is attempted as a requirement of the Reproducible Research course which is a part Data Science Specialization by John Hopkings University MOOC via Coursera Storm Data is an official publication of the National Oceanic and Atmospheric Administration (NOAA) which documents:

1. The occurrence of storms and other significant weather phenomena having sufficient intensity to cause loss of life, injuries, significant property damage, and/or disruption to commerce;
2. Rare, unusual, weather phenomena that generate media attention, such as snow flurries in South Florida or the San Diego coastal area; and
3. Other significant meteorological events, such as record maximum or minimum temperatures or precipitation that occur in connection with another event.

NCDC receives Storm Data from the National Weather Service. The National Weather service receives their information from a variety of sources, which include but are not limited to: county, state and federal emergency management officials, local law enforcement officials, skywarn spotters, NWS damage surveys, newspaper clipping services, the insurance industry and the general public.

Assignment

The basic goal of this assignment is to explore the NOAA Storm Database and answer some basic questions about severe weather events. You must use the database to answer the questions below and show the code for your entire analysis. Your analysis can consist of tables, figures, or other summaries. You may use any R package you want to support your analysis.

Questions

Your data analysis must address the following questions:

1. Across the United States, which types of events (as indicated in the `EVTYPE`EVTYPE variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

Consider writing your report as if it were to be read by a government or municipal manager who might be responsible for preparing for severe weather events and will need to prioritize resources for different types of events. However, there is no need to make any specific recommendations in your report

Preprocessing of Data

Loading libraries

Setting up preprocessing environment by loading relevant R libraries are crucial. make sure the relevant R environments has installed following packaged before hand. dplyr,tidyr, ggplot & plyr.

```
library(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
```

```
library(tidyr)
library(ggplot2)
library(plyr)
```

```
## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

Loading data

the relevant dataset is located at https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2

following script line would check wether storm dataset is already exists, if not it'll downloaded from the above mentioned link. which will prevent unnecessary data usage.

```r
#Database Loading operation
if(!exists("storm.data")) {
  #downloading operation
  if(!file.exists("StormData.csv.bz2")){
    download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2","StormData
    print("Storm database file is downloaded!")
  } else {print("Storm database file is already downloaded!")}
  storm.data <- read.csv("StormData.csv.bz2", header = TRUE)
  print("storm database loaded!")
} else {print("database already exists!")}
```

```
## [1] "Storm database file is already downloaded!"
## [1] "storm database loaded!"
```

Examination of the Data set

in the storm.data database there are **37** Columns and **902297** Rows, which can be identified using checking data dimentions.

```r
dim(storm.data)  #Check Dimentions
```

```
## [1] 902297     37
```

```r
names(storm.data) #Check headers
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

```r
str(storm.data) #Check the structure
```

```
## '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 0:00:00" .
##  $ 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 ...
```

```
head(storm.data) #Check first few lines
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE  EVTYPE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL TORNADO
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL TORNADO
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL TORNADO
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL TORNADO
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL TORNADO
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL TORNADO
##   BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END COUNTYENDN
## 1         0                                               0         NA
## 2         0                                               0         NA
## 3         0                                               0         NA
## 4         0                                               0         NA
## 5         0                                               0         NA
## 6         0                                               0         NA
##   END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES INJURIES PROPDMG
## 1         0                     14.0   100 3   0          0       15    25.0
## 2         0                      2.0   150 2   0          0        0     2.5
## 3         0                      0.1   123 2   0          0        2    25.0
## 4         0                      0.0   100 2   0          0        2     2.5
## 5         0                      0.0   150 2   0          0        2     2.5
## 6         0                      1.5   177 2   0          0        6     2.5
##   PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES LATITUDE LONGITUDE
## 1          K       0                                         3040      8812
## 2          K       0                                         3042      8755
## 3          K       0                                         3340      8742
## 4          K       0                                         3458      8626
## 5          K       0                                         3412      8642
```

```
## 6          K       0                                         3450      8748
##   LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1       3051       8806              1
## 2          0          0              2
## 3          0          0              3
## 4          0          0              4
## 5          0          0              5
## 6          0          0              6
```

```
tail(storm.data) #Check last few lines
```

```
##        STATE__           BGN_DATE    BGN_TIME TIME_ZONE COUNTY
## 902292      47 11/28/2011 0:00:00 03:00:00 PM       CST     21
## 902293      56 11/30/2011 0:00:00 10:30:00 PM       MST      7
## 902294      30 11/10/2011 0:00:00 02:48:00 PM       MST      9
## 902295       2  11/8/2011 0:00:00 02:58:00 PM       AKS    213
## 902296       2  11/9/2011 0:00:00 10:21:00 AM       AKS    202
## 902297       1 11/28/2011 0:00:00 08:00:00 PM       CST      6
##                               COUNTYNAME STATE         EVTYPE BGN_RANGE
## 902292 TNZ001>004 - 019>021 - 048>055 - 088    TN WINTER WEATHER         0
## 902293                       WYZ007 - 017    WY      HIGH WIND         0
## 902294                       MTZ009 - 010    MT      HIGH WIND         0
## 902295                            AKZ213    AK      HIGH WIND         0
## 902296                            AKZ202    AK       BLIZZARD         0
## 902297                            ALZ006    AL     HEAVY SNOW         0
##        BGN_AZI BGN_LOCATI           END_DATE    END_TIME COUNTY_END COUNTYENDN
## 902292                   11/29/2011 0:00:00 12:00:00 PM          0         NA
## 902293                   11/30/2011 0:00:00 10:30:00 PM          0         NA
## 902294                   11/10/2011 0:00:00 02:48:00 PM          0         NA
## 902295                    11/9/2011 0:00:00 01:15:00 PM          0         NA
## 902296                    11/9/2011 0:00:00 05:00:00 PM          0         NA
## 902297                   11/29/2011 0:00:00 04:00:00 AM          0         NA
##        END_RANGE END_AZI END_LOCATI LENGTH WIDTH  F MAG FATALITIES INJURIES
## 902292         0                         0     0 NA   0          0        0
## 902293         0                         0     0 NA  66          0        0
## 902294         0                         0     0 NA  52          0        0
## 902295         0                         0     0 NA  81          0        0
## 902296         0                         0     0 NA   0          0        0
## 902297         0                         0     0 NA   0          0        0
##        PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO               STATEOFFIC
## 902292       0          K       0          K MEG           TENNESSEE, West
## 902293       0          K       0          K RIW WYOMING, Central and West
## 902294       0          K       0          K TFX          MONTANA, Central
## 902295       0          K       0          K AFG          ALASKA, Northern
## 902296       0          K       0          K AFG          ALASKA, Northern
## 902297       0          K       0          K HUN            ALABAMA, North
##
## 902292 LAKE - LAKE - OBION - WEAKLEY - HENRY - DYER - GIBSON - CARROLL - LAUDERDALE - TIPTON - HAYWO
## 902293                                                                                OWL CREEK & BRID
## 902294                                                                                       NORTH ROC
## 902295
## 902296
## 902297
##        LATITUDE LONGITUDE LATITUDE_E LONGITUDE_
```

```
## 902292        0         0          0          0
## 902293        0         0          0          0
## 902294        0         0          0          0
## 902295        0         0          0          0
## 902296        0         0          0          0
## 902297        0         0          0          0
##
## 902292
## 902293
## 902294
## 902295 EPISODE NARRATIVE: A 960 mb low over the southern Aleutians at 0300AKST on the 8th intensifie
## 902296 EPISODE NARRATIVE: A 960 mb low over the southern Aleutians at 0300AKST on the 8th intensifie
## 902297                            EPISODE NARRATIVE: An intense upper level low developed on the 28th
##        REFNUM
## 902292 902292
## 902293 902293
## 902294 902294
## 902295 902295
## 902296 902296
## 902297 902297
```

Subsetting the dataset accordance with the Questions [scaling down]

the key variable used for this Assignment is - EVTYPE : e.g. Toranados, flood..

For Question 1, it refers to variable such as event type & variables related to population health. specifically

1. FATALITIES : Number of fatalities
2. INJURIES : Number of Injuries

For Question 2, variable related to types of events have the greatest economic consequences includes

1.PROPDMG : property damages

2.PROPDMGEXP: Units for Property Damage (magnitudes - K,B,M)

3.CROPDMG: Crop Damage

4.CROPDMGEXP: Units for Crop Damage (magnitudes - K,BM,B)

in order to reduce processing resources the data set will be cropped down to what is needed for analysis to solve the questions of this assignment. and the cropped dataset will be checked as below

```
storm <- select(storm.data,c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG", "CR
dim(storm) #Check Dimentions
```

```
## [1] 902297      7
```

```
names(storm) #Check headers
```

```
## [1] "EVTYPE"     "FATALITIES" "INJURIES"   "PROPDMG"    "PROPDMGEXP"
## [6] "CROPDMG"    "CROPDMGEXP"
```

```r
str(storm) #Check the structure
```

```
## 'data.frame':    902297 obs. of  7 variables:
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ 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  "" "" "" "" ...
```

```r
head(storm) #Check first few lines
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 TORNADO          0       15    25.0          K       0
## 2 TORNADO          0        0     2.5          K       0
## 3 TORNADO          0        2    25.0          K       0
## 4 TORNADO          0        2     2.5          K       0
## 5 TORNADO          0        2     2.5          K       0
## 6 TORNADO          0        6     2.5          K       0
```

```r
tail(storm) #Cheack last few line
```

```
##                EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 902292 WINTER WEATHER          0        0       0          K       0          K
## 902293      HIGH WIND          0        0       0          K       0          K
## 902294      HIGH WIND          0        0       0          K       0          K
## 902295      HIGH WIND          0        0       0          K       0          K
## 902296       BLIZZARD          0        0       0          K       0          K
## 902297     HEAVY SNOW          0        0       0          K       0          K
```

Missing Values/NA's resolution

There are 0 Missing Values /NA's in this dataset

Data Analysis

Population health dynamics related to event.

1. Fatalities

the event type has to be converted to factor variable. then all variables of EVTYPE & FATALITIES were aggregated. out of that portion top 10 were selected and arranged in decending order as following code

```r
#Analysis of Fatalities with Event type
storm$EVTYPE<-as.factor(storm$EVTYPE)
aggr.fatalites<-aggregate(FATALITIES ~ EVTYPE, data = storm,  FUN="sum") #aggregates fatalitities
top10.fatalities<-aggr.fatalites[order(-aggr.fatalites$FATALITIES), ][1:10, ] #order top 10
```

the top 10 fatality events are as following table

```
##             EVTYPE FATALITIES
## 834        TORNADO       5633
## 130 EXCESSIVE HEAT       1903
## 153    FLASH FLOOD        978
## 275           HEAT        937
## 464      LIGHTNING        816
## 856      TSTM WIND        504
## 170          FLOOD        470
## 585    RIP CURRENT        368
## 359      HIGH WIND        248
## 19       AVALANCHE        224
```

2. Injuries

variables of EVTYPE & INJURIES were aggregated. out of that portion top 10 were selected and arranged in decending order as following code

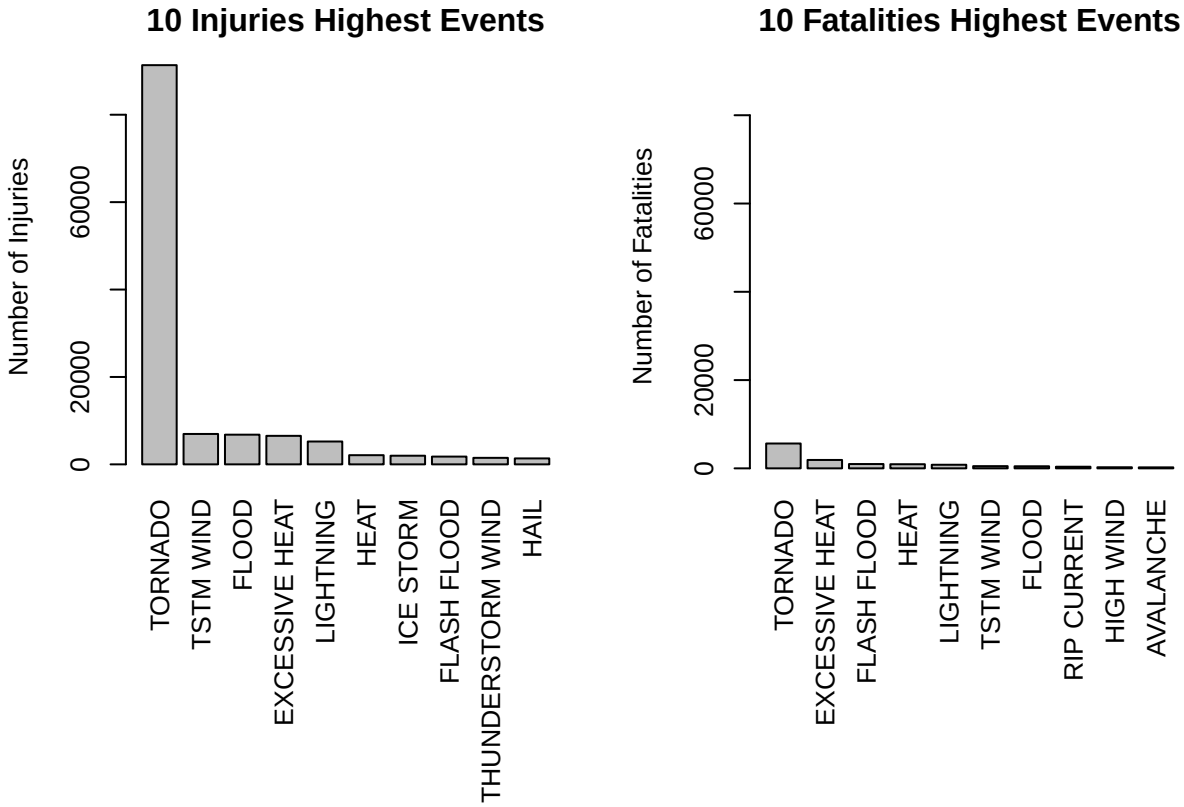
```
aggr.injuries<-aggregate(INJURIES ~ EVTYPE, data = storm,  FUN="sum") #aggregates injuries
top10.injuries<-aggr.injuries[order(-aggr.injuries$INJURIES), ][1:10, ] #order top 10
```

the top 10 Injury events are as following table

```
##               EVTYPE INJURIES
## 834           TORNADO    91346
## 856         TSTM WIND     6957
## 170             FLOOD     6789
## 130    EXCESSIVE HEAT     6525
## 464         LIGHTNING     5230
## 275              HEAT     2100
## 427         ICE STORM     1975
## 153       FLASH FLOOD     1777
## 760 THUNDERSTORM WIND     1488
## 244              HAIL     1361
```

Above mentioned both Number of Injuries and Number of fatalities against event type can be plotted in one figure as below.

```
#plotting of fatalaty vs Evtype chart
par(mfrow = c(1,2), mar = c(12, 4, 3, 2), mgp = c(3, 1, 0), cex = 0.8)
barplot(top10.injuries$INJURIES, names.arg = top10.injuries$EVTYPE, las = 3, main = "10 Injuries Highes
barplot(top10.fatalities$FATALITIES, names.arg = top10.fatalities$EVTYPE, las = 3, main = "10 Fatalitie
```




```
dev.copy(png, "event-healthplot.png", width = 480, height = 480)
```

```
## quartz_off_screen
##                 3
```

```
dev.off()
```

```
## pdf
##   2
```

Economical Consequences related to event.

on observing the data related to economical consequences which are crop damages, and property damages the values are labeled as k, m as in a seperated column of exponentional. in order to manage that, we have to turn the exponential columns of crop/property damage from character to numeric giving a measurable variable. we do mapping of values from plyr package as follows

```
#dsiplay unique values & assign its values to the same exponential
#Property Damage
unique(storm$PROPDMGEXP)
```

```
##  [1] "K" "M" ""  "B" "m" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "H" "-" "1" "8"
```

```
storm$PROPDMGEXP <- mapvalues(storm$PROPDMGEXP, from=unique(storm$PROPDMGEXP), to =c(
  10^3,# "K"
  10^6,# "M"
  1,# ""
  10^9,# "B"
  10^6,# "m"
  10^0,# "+"
  10^0,# "0"
  10^5,# "5"
  10^6,# "6"
  10^0,# "?"
  10^4,# "4"
  10^2,# "2"
  10^3,# "3"
  10^2,# "h"
  10^7,# "7"
  10^2,# "H"
  10^1,# "-"
  10^1,# "1"
  10^8# "8"
))
storm$PROPDMGEXP <- as.numeric(as.character(storm$PROPDMGEXP))
storm$PROPDMGTOTAL <- (storm$PROPDMG * storm$PROPDMGEXP)/1000000000
#Crop Damage
storm$CROPDMGEXP<-mapvalues(storm$CROPDMGEXP, from = unique(storm$CROPDMGEXP), to = c(
10^0,  # ""
10^6,  # "M"
10^3,  # "K"
10^6,  # "m"
10^9,  # "B"
10^0,  # "?"
10^0,  # "0"
10^3,  # "k"
10^2  # "2"
)
                            )
storm$CROPDMGEXP <- as.numeric(as.character(storm$CROPDMGEXP))
storm$CROPDMGTOTAL <- (storm$CROPDMG * storm$CROPDMGEXP)/1000000000
```

after clarifying the exponential its time to calculate the damages in relation to weather event type. for this we have to calculate the total number of damages per each event type after multiplying each and every damage column with exponential column.

for Property Damage, we could run the following R code which will generate the top 10 weather events causing highest property damage

```
#Calculate property damage & Display top 10 events causing highest property damage
sumPropertyDamage <- aggregate(PROPDMGTOTAL ~ EVTYPE, data = storm,  FUN="sum")
propdmg10Total <- sumPropertyDamage[order(-sumPropertyDamage$PROPDMGTOTAL), ][1:10, ]
propdmg10Total
```

```
##                EVTYPE PROPDMGTOTAL
## 170             FLOOD   144.657710
```

```
## 411 HURRICANE/TYPHOON    69.305840
## 834           TORNADO    56.947381
## 670       STORM SURGE    43.323536
## 153       FLASH FLOOD    16.822674
## 244              HAIL    15.735268
## 402         HURRICANE    11.868319
## 848    TROPICAL STORM     7.703891
## 972      WINTER STORM     6.688497
## 359         HIGH WIND     5.270046
```

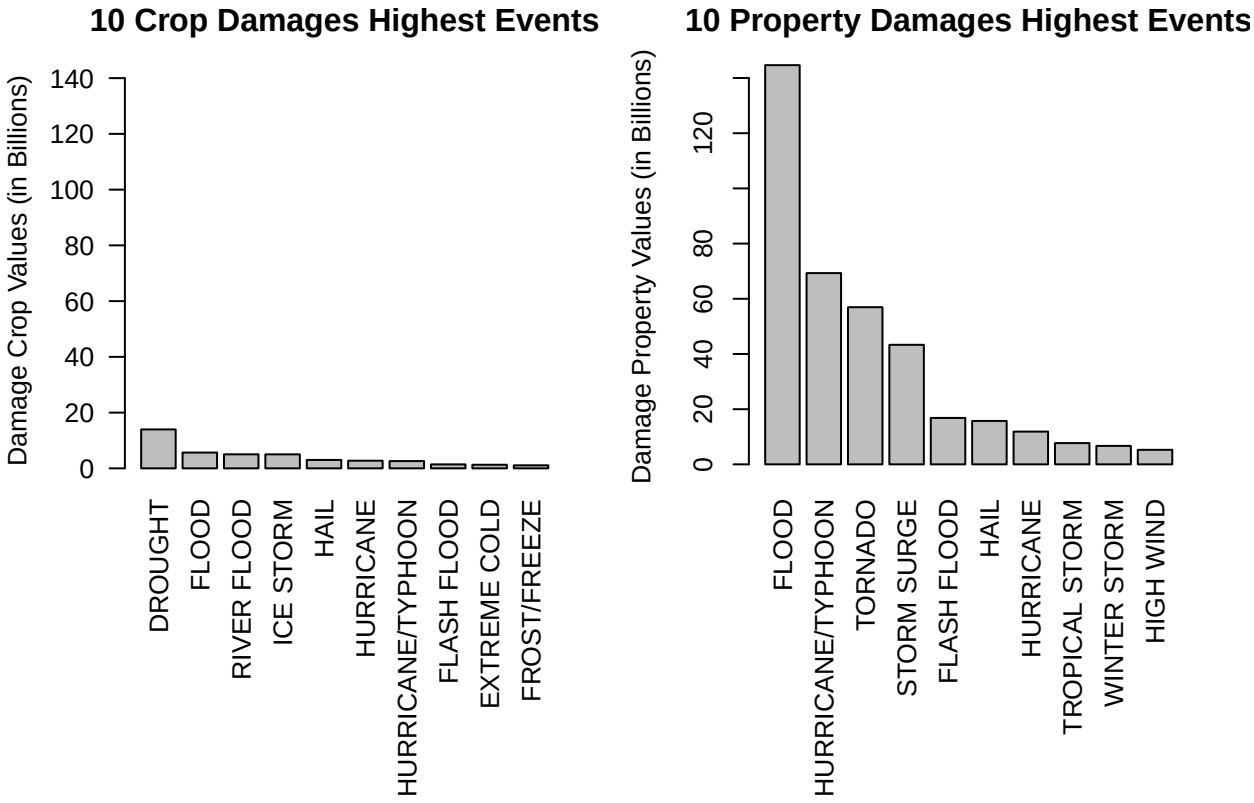
Similarly we could run the following R code which will generate the top 10 weather events causing highest crop damage

```
#Calculate crop damage & Display top 10 events causing highest crop damage
sumCropDamage <- aggregate(CROPDMGTOTAL ~ EVTYPE, data = storm,  FUN="sum")
cropdmg10Total <- sumCropDamage[order(-sumCropDamage$CROPDMGTOTAL), ][1:10, ]
cropdmg10Total
```

```
##                EVTYPE CROPDMGTOTAL
## 95            DROUGHT    13.972566
## 170             FLOOD     5.661968
## 590       RIVER FLOOD     5.029459
## 427         ICE STORM     5.022113
## 244              HAIL     3.025954
## 402         HURRICANE     2.741910
## 411 HURRICANE/TYPHOON     2.607873
## 153       FLASH FLOOD     1.421317
## 140      EXTREME COLD     1.292973
## 212      FROST/FREEZE     1.094086
```

now Finally we could create a barplot figure comparing both property and crop damage in relation to type of weather event.

```
par(mfrow = c(1,2), mar = c(12, 4, 3, 2), mgp = c(3, 1, 0), cex = 0.8)
barplot(cropdmg10Total$CROPDMGTOTAL, names.arg = cropdmg10Total$EVTYPE, las = 2, main = "10 Crop Damages
barplot(propdmg10Total$PROPDMGTOTAL, names.arg = propdmg10Total$EVTYPE, las = 3, main = "10 Property Dam
```




```
dev.copy(png, "event-economydamage.png", width = 480, height = 480)
```

```
## quartz_off_screen
##                 3
```

```
dev.off()
```

```
## pdf
##   2
```

Conclusion

based on this large data set analysis. following can be observed

in terms of population health, Injuries are far greater than fatalities in weather event. out of fatalities toranado is the leading cauase of fatality as well as injuries.

Property Damage is higher than Crop damage as economical consequences out. Flood is the leading cause of property damage while Drought is the leading cause of crop damage.