TORNADO - THE MOST HARMFUL WEATHER EVENTS ACROSS THE UNITED STATES BETWEEN 1950 AND 2011

LI Shaobai

2020/7/2

SYNOPSIS

In this study I aim to find out the most harmful weather events across the United States between 1950 and 2011, both physically and economically. The data is from National Oceanic & Atmospheric Administration (NOAA)'s National Weather Service (NWS). By calculating several sums of casualties and damages from the data, I find out that tornadoes are the most harmful weather events both physically and economically, and Texas suffers from tornadoes the most.

DATA PROCESSING

First things first, download and read in the data.

```
if (!file.exists('storm_data.bz2')){
        download.file('https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz
2','storm_data.bz2')   #download data file
}
if (!exists('storm_data')){
        storm_data=read.csv('storm_data.bz2')   #read in .bz2 file directly by read.csv
}
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
## EOF within quoted string
```

Let me take a look at the data.

```
str(storm_data)
```

```
## 'data.frame':    692288 obs. of  37 variables:
##  $ STATE__   : chr  "1.00" "1.00" "1.00" "1.00" ...
##  $ 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    : chr  "97.00" "3.00" "57.00" "89.00" ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : chr  "0.00" "0.00" "0.00" "0.00" ...
##  $ BGN_AZI   : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ COUNTY_END: chr  "0.00" "0.00" "0.00" "0.00" ...
##  $ COUNTYENDN: chr  "" "" "" "" ...
##  $ END_RANGE : chr  "0.00" "0.00" "0.00" "0.00" ...
##  $ END_AZI   : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ LENGTH    : chr  "14.00" "2.00" "0.10" "0.00" ...
##  $ WIDTH     : chr  "100.00" "150.00" "123.00" "100.00" ...
##  $ F         : chr  "3" "2" "2" "2" ...
##  $ MAG       : chr  "0.00" "0.00" "0.00" "0.00" ...
##  $ FATALITIES: chr  "0.00" "0.00" "0.00" "0.00" ...
##  $ INJURIES  : chr  "15.00" "0.00" "2.00" "2.00" ...
##  $ PROPDMG   : chr  "25.00" "2.50" "25.00" "2.50" ...
##  $ PROPDMGEXP: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : chr  "0.00" "0.00" "0.00" "0.00" ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ LATITUDE  : chr  "3040.00" "3042.00" "3340.00" "3458.00" ...
##  $ LONGITUDE : chr  "8812.00" "8755.00" "8742.00" "8626.00" ...
##  $ LATITUDE_E: chr  "3051.00" "0.00" "0.00" "0.00" ...
##  $ LONGITUDE_: chr  "8806.00" "0.00" "0.00" "0.00" ...
##  $ REMARKS   : chr  "" "" "" "" ...
##  $ REFNUM    : chr  "1.00" "2.00" "3.00" "4.00" ...
```

Some columns concerning event types and their effects-both physical and economical-interest me literally, so I just pick them out for the purpose of this study.

```
dat=storm_data[,c('EVTYPE','FATALITIES','INJURIES','PROPDMG','CROPDMG','PROPDMGEXP','CROPDMGEX
P')]
```

These column names, in my guess, mean 'EVENT TYPES', 'FATALITIES', 'INJURIES', 'PROPERTY DAMAGE', 'CROP DAMAGE', 'PROPERTY DAMAGE EXPONENTS' and 'CROP DAMAGE EXPONENTS', correspondingly.

I'm interested in the effects of the weather events not in any specific region or time period but across the whole US during the whole time period of 1950- 2011, so I just choose to overlook those columns concerning region and time and other stuff, for now.

According to the nice work entitled: How To Handle Exponent Value of PROPDMGEXP and CROPDMGEXP (https://rstudio-pubs-static.s3.amazonaws.com/58957_37b6723ee52b455990e149edde45e5b6.html), some adjustment need to be done to the 'PROPDMGEXP' and 'CROPDMGEXP' to get the actual multiplier for 'PROPDMG' and 'CROPDMG' to calculate the actual damage. The values that the post above didn't mention are simply assigned as 1 in order to keep the value of 'PROPDMG' and 'CROPDMG'.

```
dat$PROPDMGMULT=1
dat$PROPDMGMULT[which(dat$PROPDMGEXP %in% c('B','b'))]=1e9
dat$PROPDMGMULT[which(dat$PROPDMGEXP %in% c('M','m'))]=1e6
dat$PROPDMGMULT[which(dat$PROPDMGEXP %in% c('K','k'))]=1000
dat$PROPDMGMULT[which(dat$PROPDMGEXP %in% c('H','h'))]=100
dat$PROPDMGMULT[which(dat$PROPDMGEXP %in% c('0','1','2','3','4','5','6','7','8'))]=10
dat$PROPDMGMULT[which(dat$PROPDMGEXP %in% c('-','?',' ',''))]=0
dat$CROPDMGMULT=1
dat$CROPDMGMULT[which(dat$CROPDMGEXP %in% c('B','b'))]=1e9
dat$CROPDMGMULT[which(dat$CROPDMGEXP %in% c('M','m'))]=1e6
dat$CROPDMGMULT[which(dat$CROPDMGEXP %in% c('K','k'))]=1000
dat$CROPDMGMULT[which(dat$CROPDMGEXP %in% c('H','h'))]=100
dat$CROPDMGMULT[which(dat$CROPDMGEXP %in% c('0','1','2','3','4','5','6','7','8'))]=10
dat$CROPDMGMULT[which(dat$CROPDMGEXP %in% c('-','?',' ',''))]=0
```

Now that the actual damage needs to be calculated. Since characters cannot be used for calculation, some data type transformation is necessary.

```
dat[,2:5]=sapply(dat[,2:5],as.numeric)  #transform characters to numerics
```

```
## Warning in lapply(X = X, FUN = FUN, ...): 强制改变过程中产生了NA

## Warning in lapply(X = X, FUN = FUN, ...): 强制改变过程中产生了NA

## Warning in lapply(X = X, FUN = FUN, ...): 强制改变过程中产生了NA

## Warning in lapply(X = X, FUN = FUN, ...): 强制改变过程中产生了NA
```

```
dat$PROPDMGTOT=dat$PROPDMG*dat$PROPDMGMULT
dat$CROPDMGTOT=dat$CROPDMG*dat$CROPDMGMULT
```

Let me take a look at the data again.

```
summary(dat)
```

```
##     EVTYPE            FATALITIES         INJURIES          PROPDMG
##  Length:692288      Min.    :   0.00   Min.    :   0.00   Min.    : -88.31
##  Class :character   1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:   0.00
##  Mode  :character   Median :   0.00   Median :   0.00   Median :   0.00
##                     Mean   :   0.24   Mean   :   0.39   Mean   :  12.48
##                     3rd Qu.:   0.00   3rd Qu.:   0.00   3rd Qu.:   0.01
##                     Max.   :2011.00   Max.   :2011.00   Max.   :2011.00
##                     NA's   :144837    NA's   :144846    NA's   :144830
##     CROPDMG          PROPDMGEXP         CROPDMGEXP         PROPDMGMULT
##  Min.    : -87.87   Length:692288      Length:692288      Min.   :0.00e+00
##  1st Qu.:   0.00   Class :character   Class :character   1st Qu.:0.00e+00
##  Median :   0.00   Mode  :character   Mode  :character   Median :0.00e+00
##  Mean   :   1.57                                         Mean   :3.98e+04
##  3rd Qu.:   0.00                                         3rd Qu.:1.00e+00
##  Max.   :2011.00                                         Max.   :1.00e+09
##  NA's   :144838
##   CROPDMGMULT          PROPDMGTOT           CROPDMGTOT
##  Min.   :0.000e+00   Min.   :       -88   Min.   :       -88
##  1st Qu.:0.000e+00   1st Qu.:         0   1st Qu.:         0
##  Median :0.000e+00   Median :         0   Median :         0
##  Mean   :8.972e+03   Mean   :    273646   Mean   :     62523
##  3rd Qu.:0.000e+00   3rd Qu.:        10   3rd Qu.:         0
##  Max.   :1.000e+09   Max.   :5420000000   Max.   :5000000000
##                      NA's   :144830       NA's   :144838
```

It seems that the data type transformation turned a considerable part of the original characters into NAs instead of numerics, yet I see this as inevitable and don't see any feasible solution for this.

Moreover, there is at least one negative value in each of 'PROPDMGTOT' and 'CROPDMGTOT', which is meaningless. Let me see if there are more negative values in those two columns.

```
sum(dat$PROPDMGTOT<0,na.rm = T) #count negative values in 'PROPDMGTOT'
```

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

```
sum(dat$CROPDMGTOT<0,na.rm = T) #count negative values in 'CROPDMGTOT'
```

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

It seems that they are the only two negative values. Considering that the task is to search for the most harmful events and the maximum values from the data summary above are relatively big enough, I shall ignore these two negative values.

So far, the preliminary data processing is done.

DETERMINE THE MOST HARMFUL EVENTS BY CALCULATION

I decided to use sums instead of means of each event type as the index for this study, since means may be gravely affected by isolated extreme cases, while sums are more stable with reflecting the whole picture.

```
#split the data by event type, then sum 'FATILITIES', 'INJURIES', 'PROPDAMAGETOT' and 'CROPDMGTOT'
s=split(dat,dat$EVTYPE)
sums=as.data.frame(t(sapply(s,function(x) {sapply(x[,c(2,3,10,11)], function(y) sum(y,na.rm = T))})))
sums$CAS=sums$FATALITIES+sums$INJURIES  #total casualties = fatalities + injuries
sums$DMG=sums$PROPDMGTOT+sums$CROPDMGTOT  #total damage = property damage + crop damage
sums$EVTYPE=row.names(sums)     #prepare event types
```

Let me take a look at the data again.

```
summary(sums)
```

```
##    FATALITIES          INJURIES          PROPDMGTOT
##  Min.   :   0.000   Min.   :    0.00   Min.   :-8.800e+01
##  1st Qu.:   0.000   1st Qu.:    0.00   1st Qu.: 0.000e+00
##  Median :   0.000   Median :    0.00   Median : 0.000e+00
##  Mean   :   4.008   Mean   :    6.51   Mean   : 4.565e+06
##  3rd Qu.:   0.000   3rd Qu.:    0.00   3rd Qu.: 0.000e+00
##  Max.   :6024.000   Max.   :80084.00   Max.   : 4.098e+10
##    CROPDMGTOT              CAS                DMG
##  Min.   :       -88   Min.   :    0.00   Min.   :-8.800e+01
##  1st Qu.:         0   1st Qu.:    0.00   1st Qu.: 0.000e+00
##  Median :         0   Median :    0.00   Median : 0.000e+00
##  Mean   :   1042913   Mean   :   10.52   Mean   : 5.608e+06
##  3rd Qu.:         0   3rd Qu.:    0.00   3rd Qu.: 0.000e+00
##  Max.   :9860245000   Max.   :84742.00   Max.   : 4.119e+10
##     EVTYPE
##  Length:32820
##  Class :character
##  Mode  :character
##
##
##
```

Now let me determine the most harmful events.

```
sums$EVTYPE[which.max(sums$INJURIES)]   #most injuries
```

```
## [1] "TORNADO"
```

```
sums$EVTYPE[which.max(sums$FATALITIES)] #most fatalities
```

```
## [1] " 2008"
```

```
sums$EVTYPE[which.max(sums$CAS)]        #most casualties
```

```
## [1] "TORNADO"
```

```
sums$EVTYPE[which.max(sums$PROPDMGTOT)] #most property damage
```

```
## [1] "TORNADO"
```

```
sums$EVTYPE[which.max(sums$CROPDMGTOT)] #most crop damage
```

```
## [1] "DROUGHT"
```

```
sums$EVTYPE[which.max(sums$DMG)]        #most damage
```

```
## [1] "TORNADO"
```

All the results above are clear except the one of 'FATALITIES'. ' 2008' may mean that the death number peaked at the year 2008. However, let me look behind for a more meaningful answer.

```
with(sums,EVTYPE[order(FATALITIES,decreasing = T)[1:10]]) #top 10 event types by fatalities
```

```
##  [1] " 2008"
##  [2] " get named by the National Hurricane Center (NHC) as a tropical storm and/or hurrican
e. While the National Weather Service (NWS) in San Juan and the NHC closely monitored the syste
m for possible organization into a named storm"
##  [3] "TORNADO"
##  [4] " three were rated EF1"
##  [5] " along and south of Interstate 80. Accumulations of 6 to 8 inches were common in this
area. North of the interstate"
##  [6] " 2 homes were destroyed with significant portions of the homes not found.  Further nor
theast along County Road 1589"
##  [7] " a wide swath of pine and hardwood trees was found snapped at the bases or splintered
several feet off the ground"
##  [8] " an EF1 in Jefferson County.EVENT NARRATIVE: A tree was blown down onto a vehicle alon
g Brook Highland Lane.,857421.00\n1.00,4/11/2011 0:00:00,06:45:00 PM,CST,117.00,SHELBY,AL,THUND
ERSTORM WIND,2.00,SW,LANDMARK,4/11/2011 0:00:00,06:45:00 PM,0.00,,0.00,,,0.00,0.00,,50.00,0.00,
0.00,0.00,K,0.00,K,BMX,ALABAMA"
##  [9] " and 45 knots at Youngstown.EVENT NARRATIVE: ,895628.00\n45.00,8/26/2011 0:00:00,02:4
5:00 PM,EST,50.00,SCZ050,SC,HIGH SURF,0.00,,,8/26/2011 0:00:00,02:46:00 PM,0.00,,0.00,,,0.00,0.
00,,0.00,0.00,0.00,0.00,K,0.00,K,CHS,SOUTH CAROLINA"
## [10] " as rush hour was just unfolding as the tornadoes neared the Oklahoma City metro area.
EVENT NARRATIVE: ,861924.00\n30.00,5/10/2011 0:00:00,10:30:00 AM,MST,3.00,BIG HORN,MT,FLOOD,1.0
0,E,ST XAVIER,5/12/2011 0:00:00,08:00:00 AM,0.00,,21.00,N,PRYOR,0.00,0.00,,0.00,0.00,0.00,0.00,
K,0.00,K,BYZ,MONTANA"
```

It turns out that 'TORNADO' ranks third, which is the top meaningful answer.

RESULTS

It's clear that 'TORNADO' is the answer for all but crop damage, whose answer is 'DROUGHT'. However, 'TORNADO' is still the answer for total damage.

So far, I allege that tornadoes are the most harmful weather event across the US between 1950 and 2011, with respect to both population health and economy.

Let me show it with some plots.

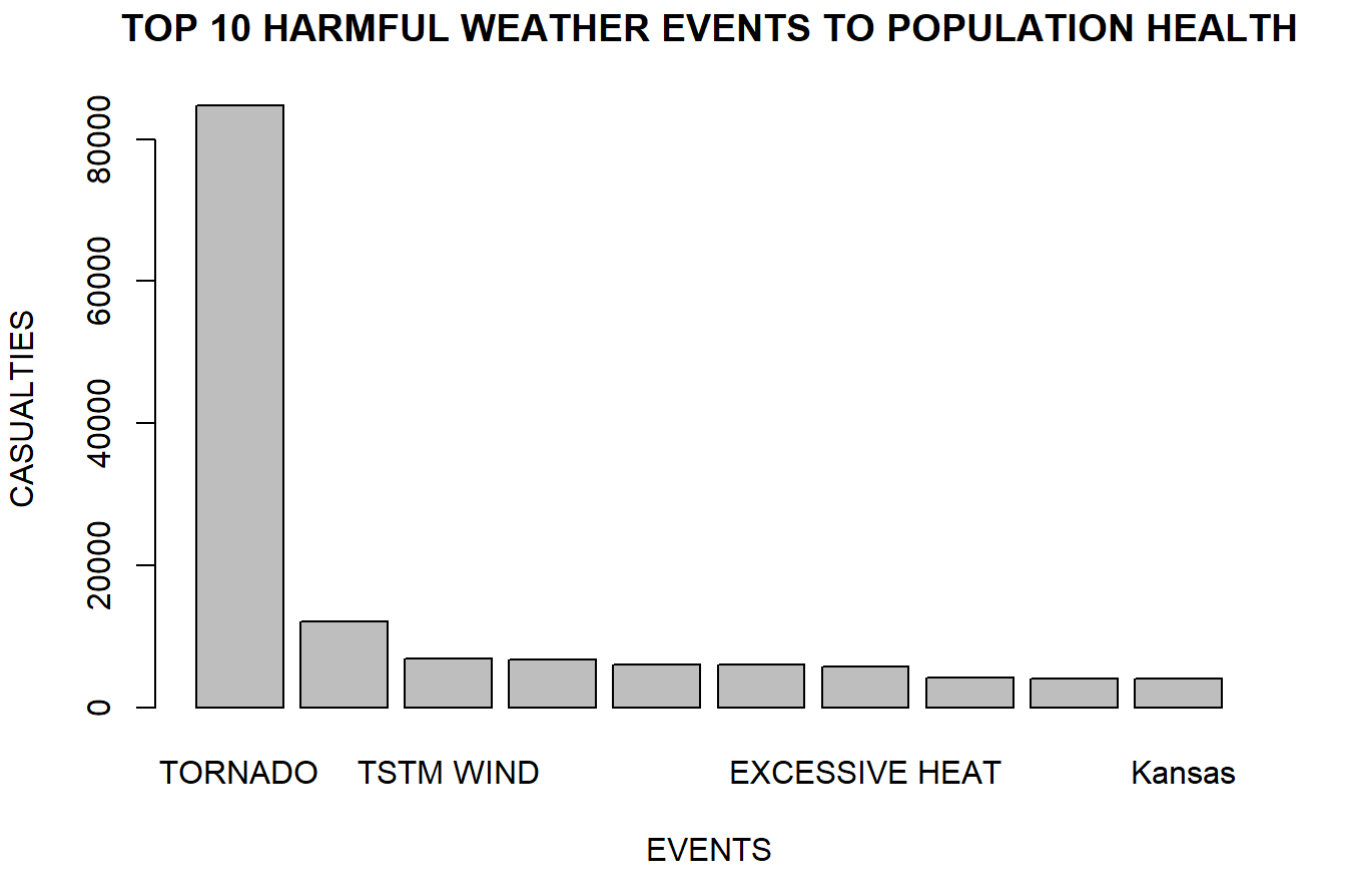
```
EVTYPECAS=with(sums,EVTYPE[order(CAS,decreasing = T)[1:10]])    #top 10 event types by casualties
EVTYPECAS
```

```
##  [1] "TORNADO"
##  [2] " 2008"
##  [3] "TSTM WIND"
##  [4] "FLOOD"
##  [5] " get named by the National Hurricane Center (NHC) as a tropical storm and/or hurricane. While the National Weather Service (NWS) in San Juan and the NHC closely monitored the system for possible organization into a named storm"
##  [6] " along and south of Interstate 80. Accumulations of 6 to 8 inches were common in this area. North of the interstate"
##  [7] "EXCESSIVE HEAT"
##  [8] "LIGHTNING"
##  [9] " three were rated EF1"
## [10] " Kansas"
```

```
TOPCAS=with(sums,CAS[order(CAS,decreasing = T)[1:10]])  #top 10 casualties
TOPCAS
```

```
##  [1] 84742 12048  6923  6757  6024  6021  5770  4190  4018  4017
```

```
barplot(height = TOPCAS,names.arg = EVTYPECAS,xlab='EVENTS',ylab='CASUALTIES',main='TOP 10 HARMFUL WEATHER EVENTS TO POPULATION HEALTH')   #plot
```



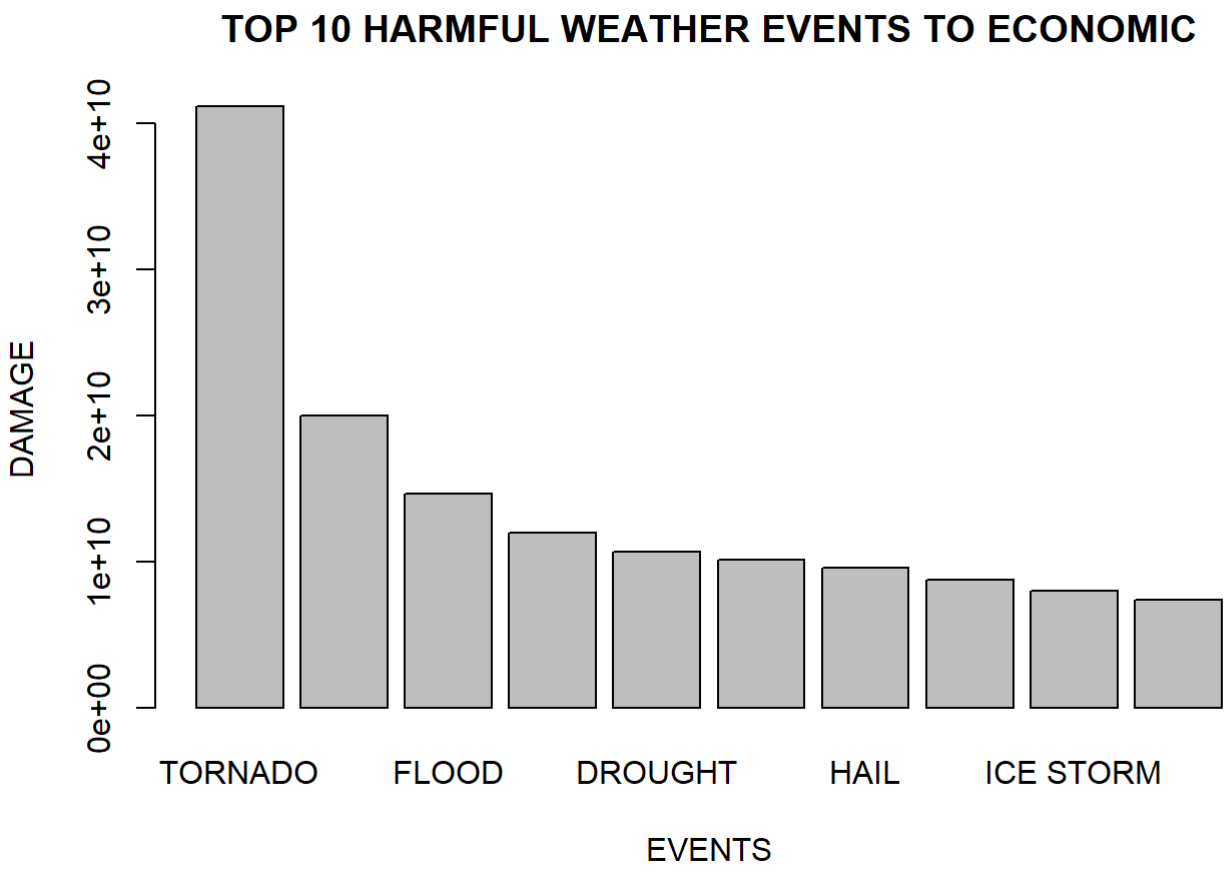
```
EVTYPEDMG=with(sums,EVTYPE[order(DMG,decreasing = T)[1:10]])    #top 10 event types by damage
EVTYPEDMG
```

```
##  [1] "TORNADO"           "HURRICANE/TYPHOON" "FLOOD"
##  [4] "HURRICANE"         "DROUGHT"           "RIVER FLOOD"
##  [7] "HAIL"              "FLASH FLOOD"       "ICE STORM"
## [10] "TROPICAL STORM"
```

```
TOPDMG=with(sums,DMG[order(DMG,decreasing = T)[1:10]])  #top 10 damage
TOPDMG
```

```
##  [1] 41193724517 19990185800 14672056020 11962119010 10705543000 10148404500
##  [7]  9568255177  8743751901  7975060010  7413537000
```

```
barplot(height = TOPDMG,names.arg = EVTYPEDMG, xlab='EVENTS',ylab='DAMAGE',main='TOP 10 HARMFUL
WEATHER EVENTS TO ECONOMIC')    #plot
```



Now it's reasonable to check which state suffers from tornadoes the most. My guess is Texas. Let me check it out.

```
ddat=storm_data[,c('STATE','EVTYPE')]   #subset the original data
tornado=subset(ddat,ddat$EVTYPE=='TORNADO')
count=as.data.frame(table(tornado$STATE),stringsAsFactors = F)  #count tornadoes by states
str(count)      #4th quick view of data
```

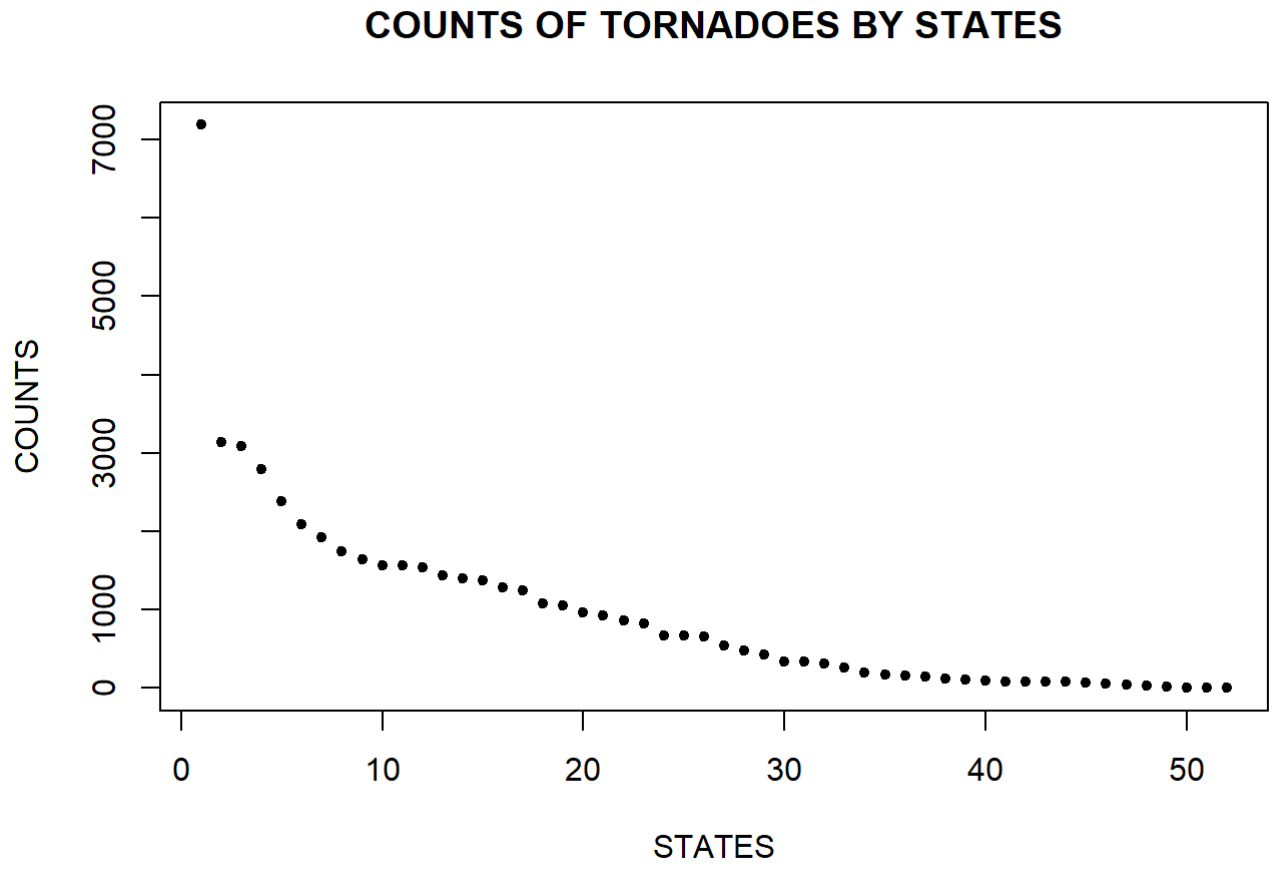
```
## 'data.frame':    52 obs. of  2 variables:
##  $ Var1: chr  "AK" "AL" "AR" "AZ" ...
##  $ Freq: int  2 1438 1541 196 316 1561 81 1 58 2786 ...
```

```
count=count[order(count$Freq,decreasing=T),]     #rank
head(count)      #check states at the top
```

	Var1 <chr>	Freq <int>
45	TX	7186
37	OK	3142
17	KS	3086
10	FL	2786
30	NE	2376
13	IA	2093

6 rows

```
plot(count$Freq,pch=20,xlab='STATES',ylab='COUNTS',main='COUNTS OF TORNADOES BY STATES')
#plot
```



Clearly, the dot at the top left is TX, which suffers from tornadoes the most. Bingo!

CONTACT

I’m on my way of accomplishing Johns Hopkins’ data science specialization and becoming a professional in data science. I’m now available to any opportunity. Please feel free to contact me for any further discussion.

Email: frankbluemoon29@icloud.com (mailto:frankbluemoon29@icloud.com)

GitHub: LI Shaobai’s GitHub (https://github.com/frank-lishaobai)