Data Processing

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : Factor w/ 16335 levels "1/1/1966 0:00:00",..: 6523 6523 4
242 11116 2224 2224 2260 383 3980 3980 ...
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683
2584 3186 242 1683 3186 3186 ...
##  $ TIME_ZONE : Factor w/ 22 levels "ADT","AKS","AST",..: 7 7 7 7 7 7 7 7
7 7 ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: Factor w/ 29601 levels "","5NM E OF MACKINAC BRIDGE TO PR
ESQUE ISLE LT MI",..: 13513 1873 4598 10592 4372 10094 1973 23873 24418 459
8 ...
##  $ STATE     : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834
834 834 834 834 834 834 834 834 ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : Factor w/ 35 levels "","  N"," NW",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ BGN_LOCATI: Factor w/ 54429 levels "","- 1 N Albion",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ END_DATE  : Factor w/ 6663 levels "","1/1/1993 0:00:00",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ END_TIME  : Factor w/ 3647 levels "","  0900CST",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ 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   : Factor w/ 24 levels "","E","ENE","ESE",..: 1 1 1 1 1 1 1
1 1 1 ...
##  $ END_LOCATI: Factor w/ 34506 levels "","- .5 NNW",..: 1 1 1 1 1 1 1 1
1 1 ...
##  $ 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: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17
17 17 17 17 ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 1 1 1 1 1 1
...
##  $ WFO       : Factor w/ 542 levels "","  CI","$AC",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ STATEOFFIC: Factor w/ 250 levels "","ALABAMA, Central",..: 1 1 1 1 1
1 1 1 1 1 ...
##  $ ZONENAMES : Factor w/ 25112 levels "","
"| __truncated__,..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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   : Factor w/ 436781 levels "","-2 at Deer Park\n",..: 1 1 1
1 1 1 1 1 1 1 ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

```
head(data$INJURIES,5)
```

```
## [1] 15  0  2  2  2
```

```
head(data$FATALITIES,5)
```

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

```
head(data$PROPDMG,5)
```

```
## [1] 25.0  2.5 25.0  2.5  2.5
```

```
head(data$PROPDMGEXP,5)
```

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

```
head(data$CROPDMG,5)
```

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

```
head(data$CROPDMGEXP,5)
```

```
## [1]
## Levels:  ? 0 2 B k K m M
```

```
# Recode the PROPDMGEXP into appropriate 'multipliers'
data$PROPEXP[data$PROPDMGEXP == "H"] <- 100          #H-> Hundreds
data$PROPEXP[data$PROPDMGEXP == "h"] <- 100
data$PROPEXP[data$PROPDMGEXP == "K"] <- 1000         #K-> Thousands
data$PROPEXP[data$PROPDMGEXP == "M"] <- 1e+06        #M-> Millions
data$PROPEXP[data$PROPDMGEXP == "m"] <- 1e+06
data$PROPEXP[data$PROPDMGEXP == "B"] <- 1e+09        #B-> Billions
```

```r
data$PROPEXP[data$PROPDMGEXP == ""] <- 1
data$PROPEXP[data$PROPDMGEXP == "0"] <- 1
data$PROPEXP[data$PROPDMGEXP == "1"] <- 10
data$PROPEXP[data$PROPDMGEXP == "2"] <- 100
data$PROPEXP[data$PROPDMGEXP == "3"] <- 1000
data$PROPEXP[data$PROPDMGEXP == "4"] <- 10000
data$PROPEXP[data$PROPDMGEXP == "5"] <- 1e+05
data$PROPEXP[data$PROPDMGEXP == "6"] <- 1e+06
data$PROPEXP[data$PROPDMGEXP == "7"] <- 1e+07
data$PROPEXP[data$PROPDMGEXP == "8"] <- 1e+08
# Invalid values
data$PROPEXP[data$PROPDMGEXP == "+"] <- 0
data$PROPEXP[data$PROPDMGEXP == "-"] <- 0
data$PROPEXP[data$PROPDMGEXP == "?"] <- 0

#Calculate for the PROPERTY  DAMAGEVALUE: Whole number x Multiplier
data$propvalue <- data$PROPDMG * data$PROPEXP

# Recode the CROPDMGEXP into appropriate 'multipliers'
data$CROPEXP[data$CROPDMGEXP == "K"] <- 1000
data$CROPEXP[data$CROPDMGEXP == "k"] <- 1000
data$CROPEXP[data$CROPDMGEXP == "M"] <- 1e+06
data$CROPEXP[data$CROPDMGEXP == "m"] <- 1e+06
data$CROPEXP[data$CROPDMGEXP == "B"] <- 1e+09
data$CROPEXP[data$CROPDMGEXP == "0"] <- 1
data$CROPEXP[data$CROPDMGEXP == "2"] <- 100
data$CROPEXP[data$CROPDMGEXP == ""] <- 1
# Invalid values
data$CROPEXP[data$CROPDMGEXP == "?"] <- 0
#Calculate for the CROP DAMAGEVALUE: Whole number x Multiplier
data$cropvalue <- data$CROPDMG * data$CROPEXP
```

Aggregate the Data.

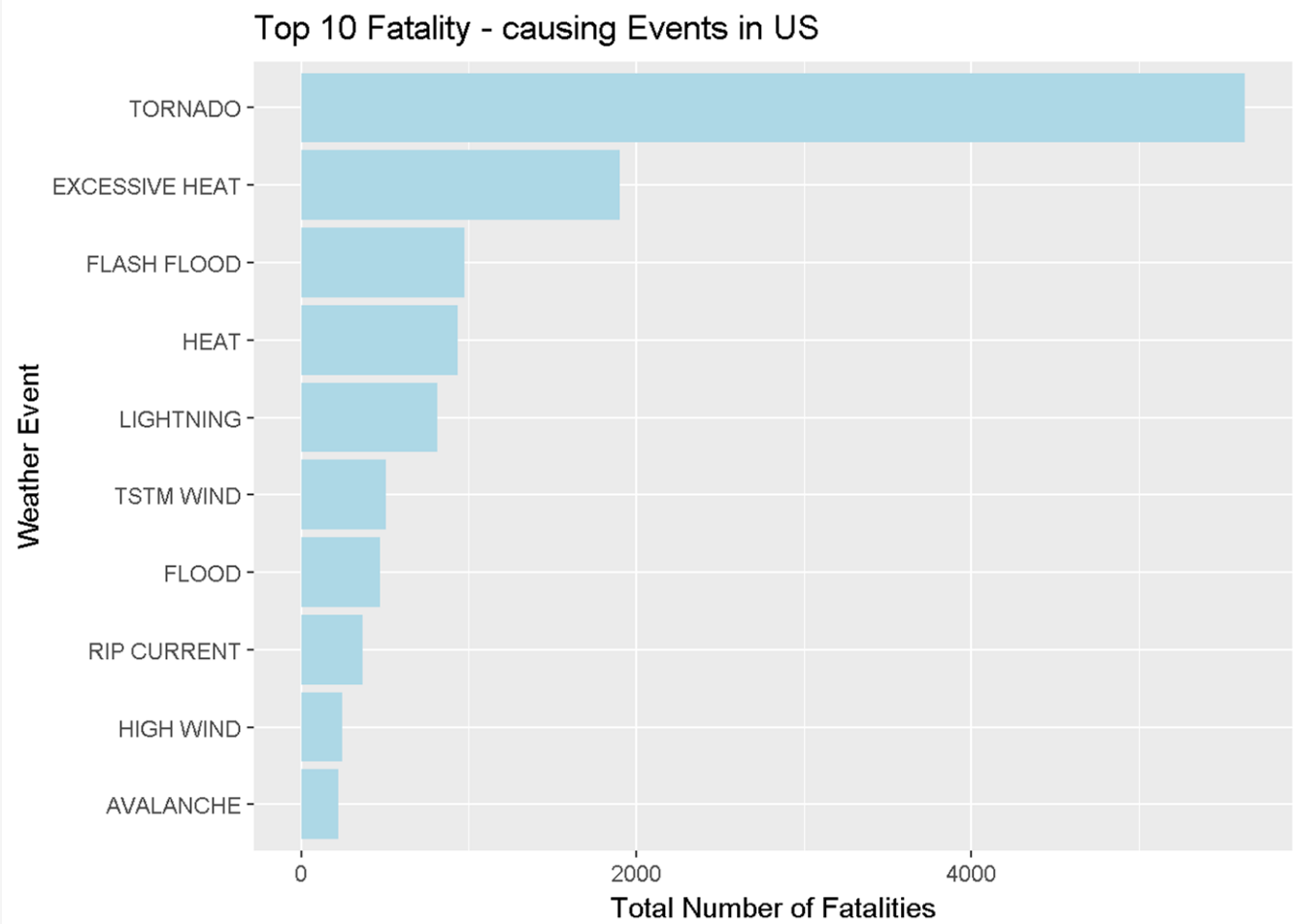
```r
#Get the total number (SUM) of ijuries and fatalities by event type.
dfatal<-aggregate(FATALITIES ~ EVTYPE, data=data, sum)
dinj<-aggregate(INJURIES ~ EVTYPE, data=data, sum)
dprop<-aggregate(propvalue ~ EVTYPE, data=data, sum)
```

```
dcrop<-aggregate(cropvalue ~ EVTYPE, data=data, sum)


#SORT THE DATA INTO DECREASING ORDER
dfatalsort<-dfatal[order(dfatal$FATALITIES,decreasing = T),]
dinjsort<-dinj[order(dinj$INJURIES,decreasing = T),]
dpropsort<-dprop[order(dprop$propvalue,decreasing = T),]
dcropsort<-dcrop[order(dcrop$cropvalue,decreasing = T),]


forgraph1<-dfatalsort[1:10,]
forgraph2<-dinjsort[1:10,]
forgraph3<-dpropsort[1:10,]
forgraph4<-dcropsort[1:10,]
forgraph3$propvalue2<-forgraph3$propvalue/(10^9)
forgraph4$cropvalue2<-forgraph4$cropvalue/(10^9)
```

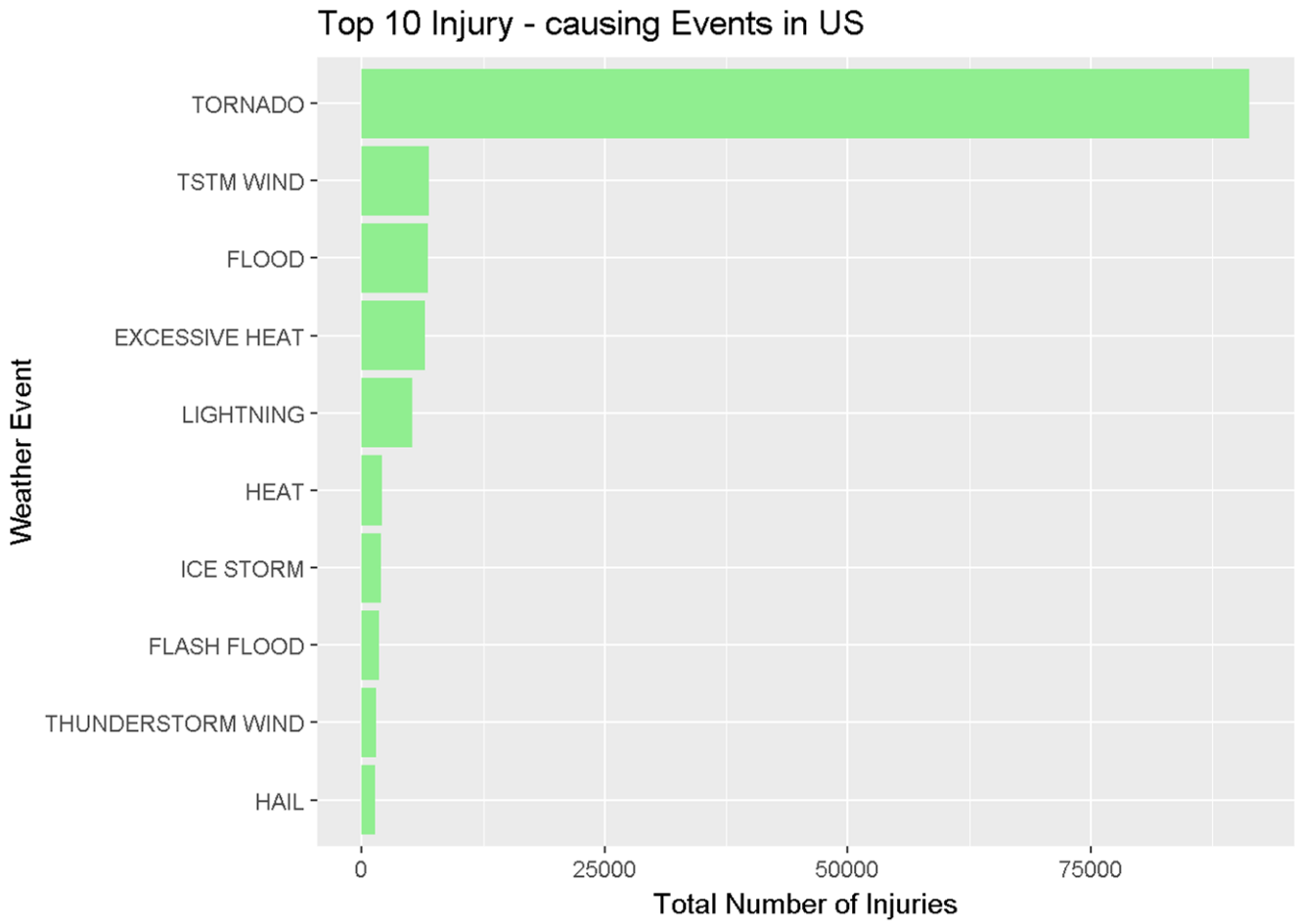
IMPACT TO POPULATION HEALTH

Show the top 10 events in terms of number of fatalities, and injuries.

```
graph1 <- ggplot(data=forgraph1, aes(x=reorder(EVTYPE, FATALITIES), y=FATAL
ITIES)) +
  geom_bar(fill="lightblue",stat="identity")  + coord_flip() +
  labs(list(title="Top 10 Fatality - causing Events in US",x="Weather Event
", y="Total Number of Fatalities"))


graph2 <- ggplot(data=forgraph2, aes(x=reorder(EVTYPE, INJURIES), y=INJURIE
S)) +
  geom_bar(fill="lightgreen",stat="identity")  + coord_flip() +
  labs(list(title="Top 10 Injury - causing Events in US",x="Weather Event",
y="Total Number of Injuries"))


graph1
```


Top 10 Fatality - causing Events in US
TORNADO
EXCESSIVE HEAT
FLASH FLOOD
HEAT
LIGHTNING
TSTM WIND
FLOOD
RIP CURRENT
HIGH WIND
AVALANCHE
Weather Event
0
2000
4000
Total Number of Fatalities

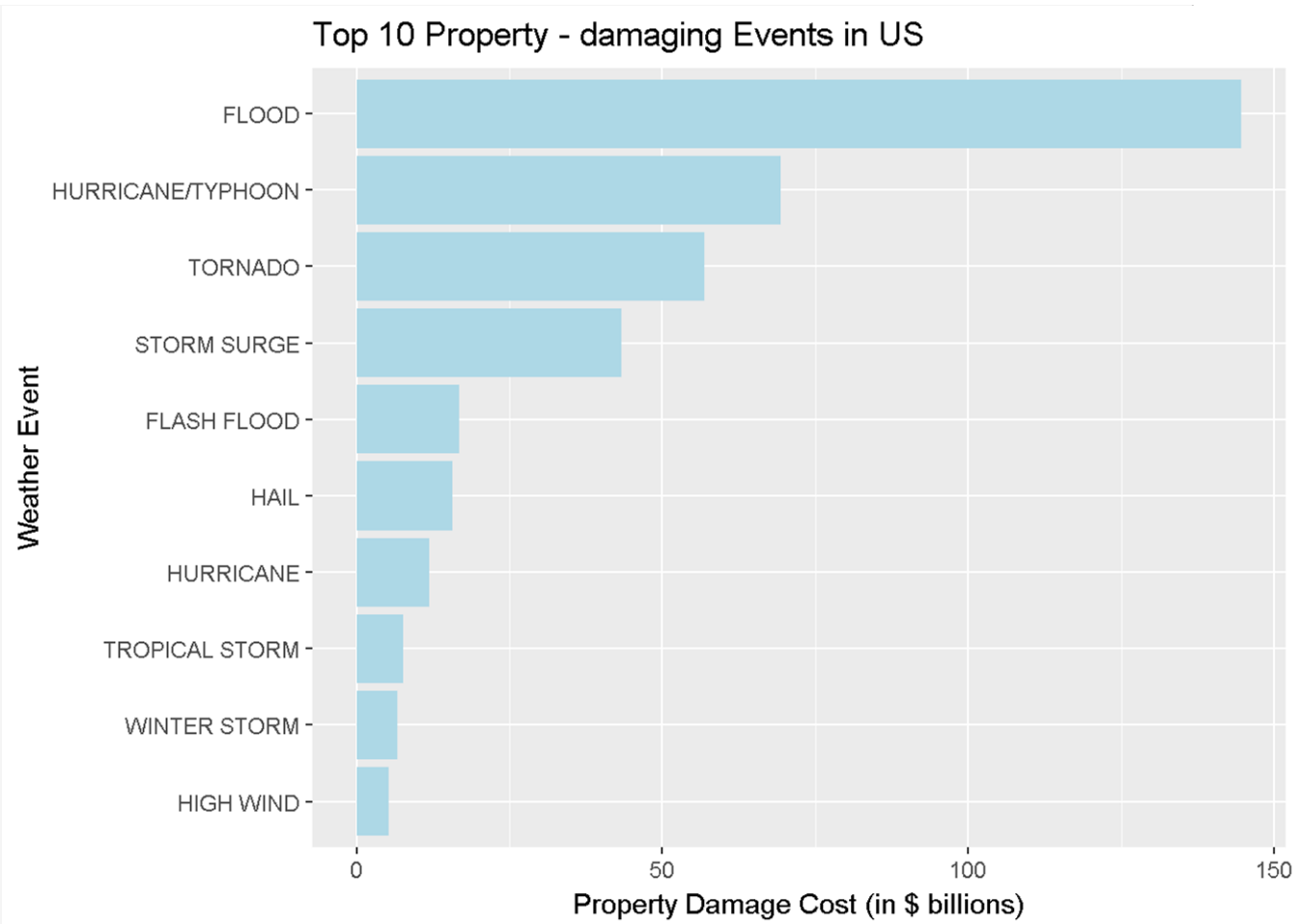
```
graph2
```



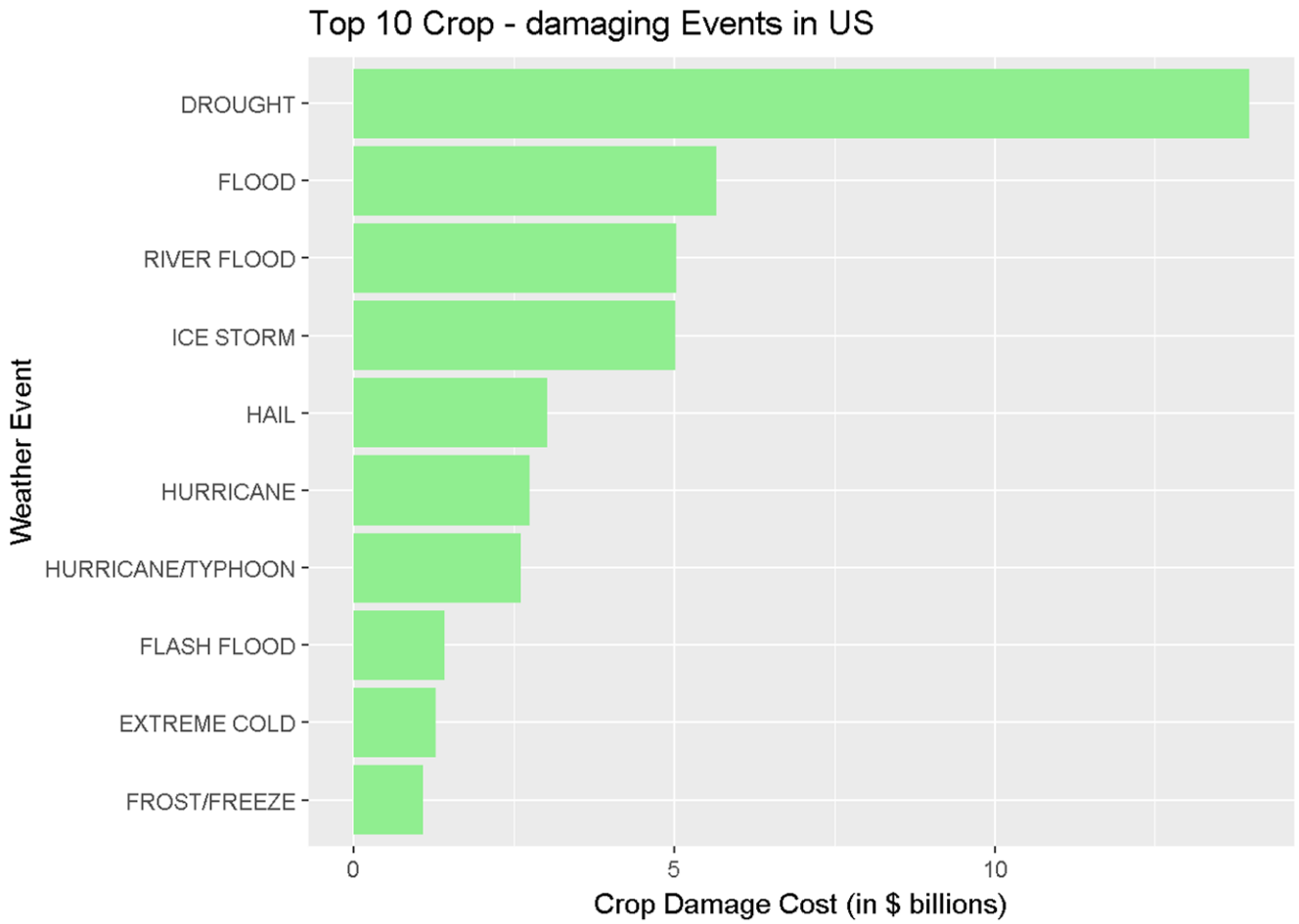
IMPACT TO ECONOMY

Show the top 10 events in terms of damage to properties and crops.

```
graph3 <- ggplot(data=forgraph3, aes(x=reorder(EVTYPE, propvalue2), y=propv
alue2)) +

  geom_bar(fill="lightblue",stat="identity")  + coord_flip() +

  labs(list(title="Top 10 Property - damaging Events in US",x="Weather Even
t", y="Property Damage Cost (in $ billions)"))


graph4 <- ggplot(data=forgraph4, aes(x=reorder(EVTYPE,cropvalue2), y=cropva
lue2)) +

  geom_bar(fill="lightgreen",stat="identity")  + coord_flip() +

  labs(list(title="Top 10 Crop - damaging Events in US",x="Weather Event",
y="Crop Damage Cost (in $ billions)"))


graph3
```


Top 10 Property - damaging Events in US
FLOOD
HURRICANE/TYPHOON
TORNADO
STORM SURGE
FLASH FLOOD
HAIL
HURRICANE
TROPICAL STORM
WINTER STORM
HIGH WIND
Weather Event
0
50
100
150
Property Damage Cost (in $ billions)

```
graph4
```



CONCLUSION

In terms of number of injuries and fatalities, TORNADO has the highest damage to people's health.

On the other hand, FLOOD has the highest property damage cost, and DROUGHT has the highest crop damage cost.