NOAA-Storm-Analysis

Synopsis

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

The analysis undertaken addresses the following questions:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

In this analysis, initially, the total injuries and fatalities are calculated for each event type. Then, the top 10 most harmful events are shortlisted. Similarly, property damages and crop damages are computed for each event type and the top 10 events with the greatest economic consequences are shortlisted. Finally, the results are presented as plots and figures.

Import Packages

```
library(dplyr)        # for data manipulation
library(ggplot2)      # for data visulization
library(grid)         # for grid graphics
library(scales)       # for scale functions
library(gtable)       # for grobs manipulation
```

Data Processing

Check and download the **Storm data** into the `data` folder.

```
# Check to see if the directory exists
if(!file.exists("./data")) {dir.create("data")}

## Check and download the data if not yet downloaded
if(!file.exists("./data/Storm_data.csv.bz2")) {
  url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
  download.file(url, destfile = "./data/Storm_data.csv.bz2")
}
```

Load the **Storm data** into a dataframe.

```
df <- read.csv("./data/Storm_data.csv.bz2")
```

Look at the first few rows of the dataset.

```
head(df)
```

```
##   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
```

Extract only the **required** columns into a new dataframe.

```
df_req <- df %>%
          select(c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPD
MG", "CROPDMGEXP"))
```

Look at the structure of the resulting dataset.

```
str(df_req)
```

```
## 'data.frame':    902297 obs. of  7 variables:
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834
834 834 834 834 834 ...
##  $ 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 ...
```

Look at the summary for the resulting dataset.

```
summary(df_req)
```

```
##                EVTYPE         FATALITIES          INJURIES
##  HAIL             :288661   Min.   :  0.0000   Min.   :   0.0000
##  TSTM WIND        :219940   1st Qu.:  0.0000   1st Qu.:   0.0000
##  THUNDERSTORM WIND: 82563   Median :  0.0000   Median :   0.0000
##  TORNADO          : 60652   Mean   :  0.0168   Mean   :   0.1557
##  FLASH FLOOD      : 54277   3rd Qu.:  0.0000   3rd Qu.:   0.0000
##  FLOOD            : 25326   Max.   :583.0000   Max.   :1700.0000
##  (Other)          :170878
##     PROPDMG          PROPDMGEXP       CROPDMG          CROPDMGEXP
##  Min.   :   0.00          :465934   Min.   :  0.000          :618413
##  1st Qu.:   0.00   K      :424665   1st Qu.:  0.000   K      :281832
##  Median :   0.00   M      : 11330   Median :  0.000   M      :  1994
##  Mean   :  12.06   0      :   216   Mean   :  1.527   k      :    21
##  3rd Qu.:   0.50   B      :    40   3rd Qu.:  0.000   0      :    19
##  Max.   :5000.00   5      :    28   Max.   :990.000   B      :     9
##                    (Other):    84                     (Other):     9
```

Tabulate the unique values in **`PROPDMGEXP`** column along with their respective counts.

```
table(df_req$PROPDMGEXP)
```

```
##
##             -      ?      +      0      1      2      3      4      5      6
## 465934      1      8      5    216     25     13      4      4     28      4
##      7      8      B      h      H      K      m      M
##      5      1     40      1      6 424665      7  11330
```

Tabulate the unique values in **`CROPDMGEXP`** column along with their respective counts.

```
table(df_req$CROPDMGEXP)
```

```
##
##             ?      0      2      B      k      K      m      M
## 618413      7     19      1      9     21 281832      1   1994
```

From above, it can be observed that there are some erroneous values in the **`PROPDMGEXP`** and **`CROPDMGEXP`** columns such as **"?"**, **"+"**, etc. Since, their proportions are low, let us drop these observations.

```
df_req <- df_req[!(df_req$PROPDMGEXP %in% c("-", "?", "+", "h", "H") | df_req$CROPDMGE
XP == "?"), ]
```

Replace the **alphabetical exponents** with their respective **numerical equivalent values**, such as **'M'** by **'6'** (i.e. *Million*), etc.

```
# Add the new factor levels
levels(df_req$PROPDMGEXP) <- c(levels(df_req$PROPDMGEXP), '0', '3', '6', '9')
levels(df_req$CROPDMGEXP) <- c(levels(df_req$CROPDMGEXP), '0', '3', '6', '9')
```

```
# Replace the aplhabetical exponents from 'PROPDMGEXP'
df_req$PROPDMGEXP[(df_req$PROPDMGEXP %in% c("B", "b"))] <- '9'
df_req$PROPDMGEXP[(df_req$PROPDMGEXP %in% c("M", "m"))] <- '6'
df_req$PROPDMGEXP[(df_req$PROPDMGEXP %in% c("K", "k"))] <- '3'
```

```
# Replace the aplhabetical exponents from 'CROPDMGEXP'
df_req$CROPDMGEXP[(df_req$CROPDMGEXP %in% c("B", "b"))] <- '9'
df_req$CROPDMGEXP[(df_req$CROPDMGEXP %in% c("M", "m"))] <- '6'
df_req$CROPDMGEXP[(df_req$CROPDMGEXP %in% c("K", "k"))] <- '3'
```

Also, replace the **'null'** factors by **'0'**.

```
# Replace the 'null' exponent from 'PROPDMGEXP'
df_req$PROPDMGEXP[df_req$PROPDMGEXP == ""] <- '0'

# Replace the 'null' exponent from 'CROPDMGEXP'
df_req$CROPDMGEXP[df_req$CROPDMGEXP == ""] <- '0'
```

Convert the **'factor'** variables to **'numeric'** variables.

```
# Convert 'PROPDMGEXP' to 'numeric'
df_req$PROPDMGEXP <- as.numeric(as.character(df_req$PROPDMGEXP))

# Convert 'CROPDMGEXP' to 'numeric'
df_req$CROPDMGEXP <- as.numeric(as.character(df_req$CROPDMGEXP))
```

Again, tabulate the unique values in **`PROPDMGEXP`** column along with their respective counts.

```
table(df_req$PROPDMGEXP)
```

```
##
##      0      1      2      3      4      5      6      7      8      9
## 466148     25     13 424665      4     28  11340      5      1     40
```

Similarly, tabulate the unique values in **`CROPDMGEXP`** column along with their respective counts.

```
table(df_req$CROPDMGEXP)
```

```
## 
##      0      2      3      6      9 
## 618411      1 281853   1995      9
```

Now, calculate the total **Property Damage** and **Crop Damage** and store them into the variables `TOTPROPDMG` and `TOTCROPDMG`, respectively.

```
# Total 'Property Damage'
df_req$TOTPROPDMG <- df_req$PROPDMG * (10 ^ df_req$PROPDMGEXP)

# Total 'Crop Damage'
df_req$TOTCROPDMG <- df_req$CROPDMG * (10 ^ df_req$CROPDMGEXP)
```

Calculate the **Total Damage** (i.e. sum of *Property Damage* and *Crop Damage*) and store it into the variable `TOTDMG`.

```
df_req$TOTDMG <- df_req$TOTPROPDMG + df_req$TOTCROPDMG
```

Lastly, convert the `EVTYPE` to **character** variable and convert to **upper** case.

```
df_req$EVTYPE <- toupper(as.character(df_req$EVTYPE))
```

Analysis

Look at the first few rows of the resulting dataset.

```
head(df_req)
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP TOTPROPDMG
## 1 TORNADO          0       15    25.0          3       0          0      25000
## 2 TORNADO          0        0     2.5          3       0          0       2500
## 3 TORNADO          0        2    25.0          3       0          0      25000
## 4 TORNADO          0        2     2.5          3       0          0       2500
## 5 TORNADO          0        2     2.5          3       0          0       2500
## 6 TORNADO          0        6     2.5          3       0          0       2500
##   TOTCROPDMG TOTDMG
## 1          0  25000
## 2          0   2500
## 3          0  25000
## 4          0   2500
## 5          0   2500
## 6          0   2500
```

Group the dataset by the event type `EVTYPE` and calculate the **total** `FATALITIES`, `INJURIES`, `TOTPROPDMG`, `TOTCROPDMG` and `TOTDMG` for each event type. Also, store the results in a new dataframe.

```
df_res <- df_req %>%
          group_by(EVTYPE) %>%
          summarise(across(c("FATALITIES", "INJURIES", "TOTPROPDMG", "TOTCROPDMG", "TOTDMG"), sum))
```

Extract the top 10 most harmful events with respect to **`FATALITIES`** .

```
fatalities10 <- df_res %>% arrange(desc(FATALITIES)) %>% head(n=10)
```

Extract the top 10 most harmful events with respect to **`INJURIES`** .

```
injuries10 <- df_res %>% arrange(desc(INJURIES)) %>% head(n=10)
```

Extract the top 10 events with greatest economic consequences with respect to **`TOTPROPDMG`** .

```
propdmg10 <- df_res %>% arrange(desc(TOTPROPDMG)) %>% head(n=10)
```

Extract the top 10 events with greatest economic consequences with respect to **`TOTCROPDMG`** .

```
cropdmg10 <- df_res %>% arrange(desc(TOTCROPDMG)) %>% head(n=10)
```

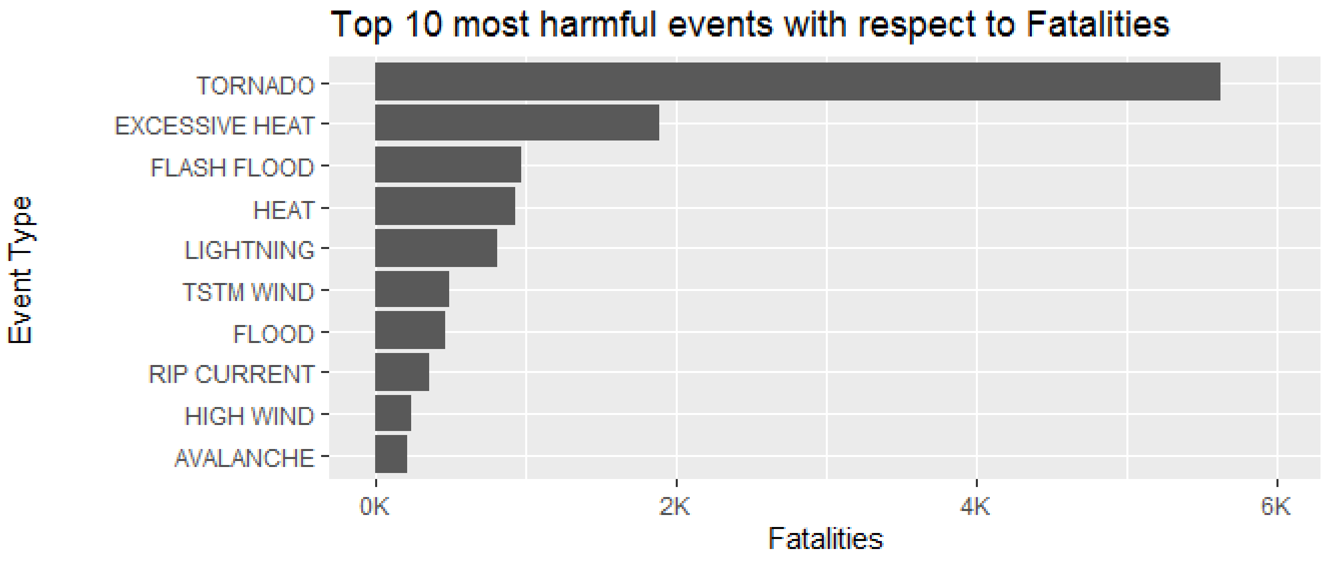
Extract the top 10 events with greatest economic consequences with respect to **`TOTDMG`** .

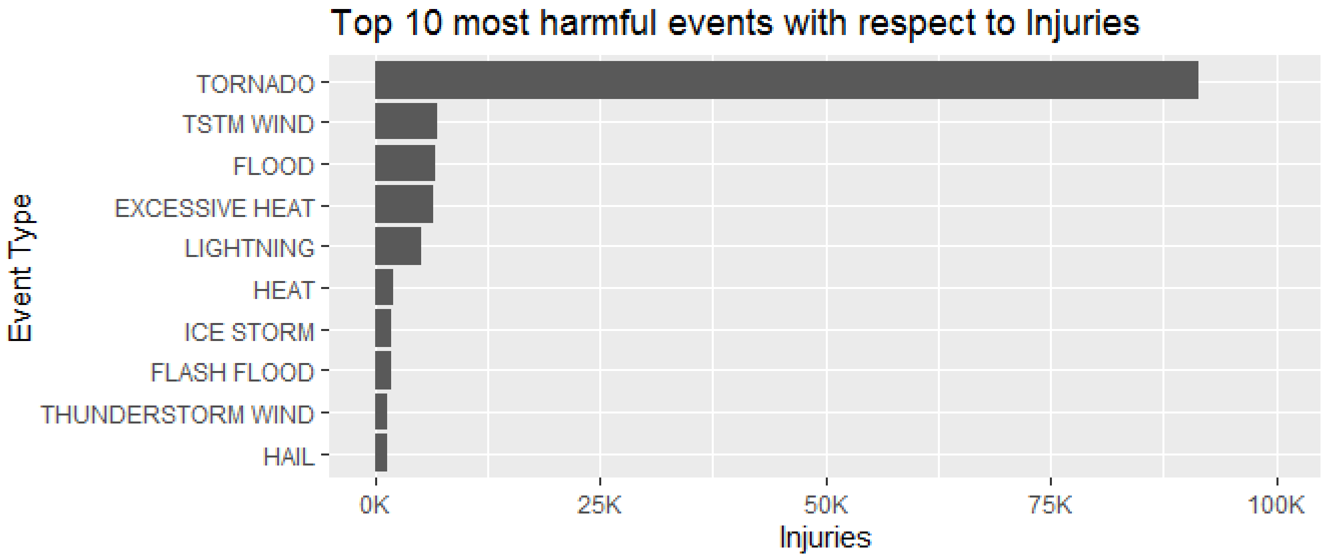
```
totdmg10 <- df_res %>% arrange(desc(TOTDMG)) %>% head(n=10)
```

Results

Look at the top 10 most harmful events with respect to population health.

```
g1 <- fatalities10 %>%
      ggplot(aes(FATALITIES, reorder(EVTYPE, FATALITIES))) +
      geom_col() +
      labs(title = "Top 10 most harmful events with respect to Fatalities", x = "Fatal
ities", y = "Event Type") +
      scale_x_continuous(limits = c(0, 6000),
                         labels = label_number(accuracy = 1, scale = 1/1000, suffix =
"K"))

g2 <- injuries10 %>%
      ggplot(aes(INJURIES, reorder(EVTYPE, INJURIES))) +
      geom_col() +
      labs(title = "Top 10 most harmful events with respect to Injuries", x = "Injurie
s", y = "Event Type") +
      scale_x_continuous(limits = c(0, 100000),
                         labels = label_number(accuracy = 1, scale = 1/1000, suffix =
"K"))

p1 <- ggplotGrob(g1)
p2 <- ggplotGrob(g2)
pl1 <- rbind(p1, p2, size = "first")
pl1$widths <- unit.pmax(p1$widths, p2$widths)
grid.newpage()
grid.draw(pl1)
```

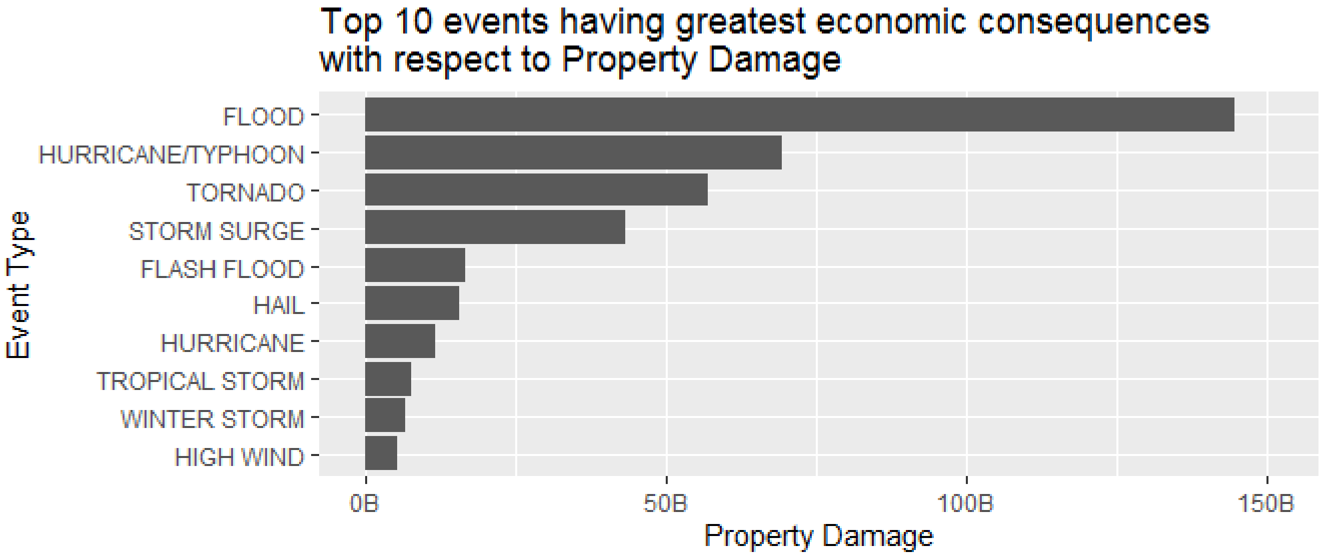


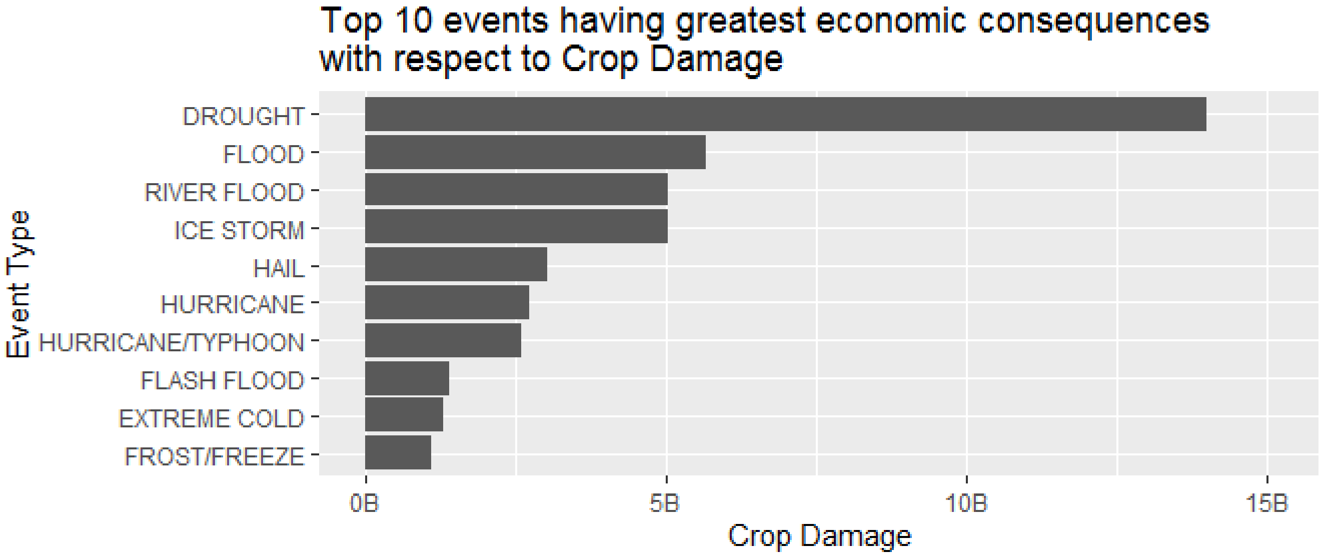


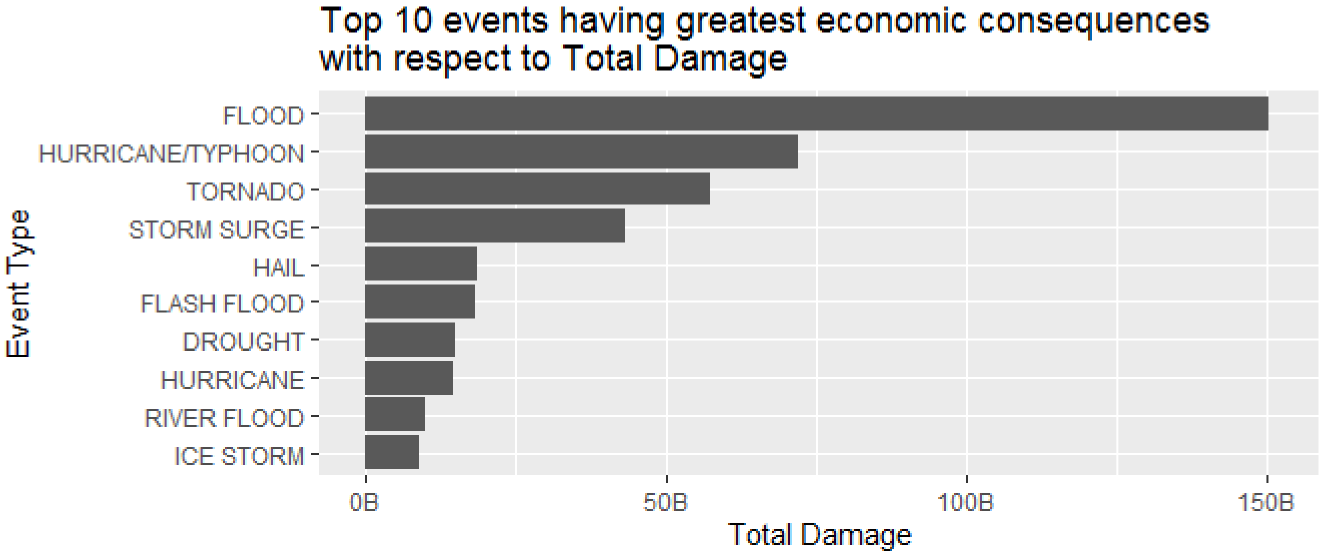
- From above, it is observed that with respect to ***Fatalities***, the most harmful event is the **Tornado** followed by **Excessive Heat** and **Flash Flood**.
- Again, with respect to ***Injuries***, the most harmful event is the **Tornado**. But, here, it is followed by **Tstm Wind** and then **Flood**. Whereas, **Excessive Heat** has dropped down to the **4th** place and **Flash Flood** down to the **8th** place.
- There is an overlap of **7** events in the two lists.
- The total number of ***Injuries*** is much higher than the number of ***Fatalities***.

Now, look at the top 10 events having the greatest economic consequences.

```
g3 <- propdmg10 %>%
      ggplot(aes(TOTPROPDMG, reorder(EVTYPE, TOTPROPDMG))) +
      geom_col() +
      labs(title = "Top 10 events having greatest economic consequences\nwith respect
 to Property Damage",
           x = "Property Damage", y = "Event Type") +
      scale_x_continuous(limits = c(0, 151000000000),
                         labels = label_number(accuracy = 1, scale = 1/1000000000, suf
fix = "B"))

g4 <- cropdmg10 %>%
      ggplot(aes(TOTCROPDMG, reorder(EVTYPE, TOTCROPDMG))) +
      geom_col() +
      labs(title = "Top 10 events having greatest economic consequences\nwith respect
 to Crop Damage",
           x = "Crop Damage", y = "Event Type") +
      scale_x_continuous(limits = c(0, 15100000000),
                         labels = label_number(accuracy = 1, scale = 1/1000000000, suf
fix = "B"))

g5 <- totdmg10 %>%
      ggplot(aes(TOTDMG, reorder(EVTYPE, TOTDMG))) +
      geom_col() +
      labs(title = "Top 10 events having greatest economic consequences\nwith respect
 to Total Damage",
           x = "Total Damage", y = "Event Type") +
      scale_x_continuous(limits = c(0, 151000000000),
                         labels = label_number(accuracy = 1, scale = 1/1000000000, suf
fix = "B"))

p3 <- ggplotGrob(g3)
p4 <- ggplotGrob(g4)
p5 <- ggplotGrob(g5)
pl2 <- rbind(p3, p4, p5, size = "first")
pl2$widths <- unit.pmax(p3$widths, p4$widths, p5$widths)
grid.newpage()
grid.draw(pl2)
```


Top 10 events having greatest economic consequences
with respect to Property Damage


Top 10 events having greatest economic consequences
with respect to Crop Damage


Top 10 events having greatest economic consequences
with respect to Total Damage

- From above, it can be observed that the two lists of top 10 events having greatest economic consequences with respect to the ***Property Damage*** and ***Crop Damage*** are considerably different.

- Since, the scale of **Property Damage** is much higher than that of **Crop Damage**, there is good overlap between the lists of top 10 events with respect to ***Property Damage*** and ***Total Damage***, such as **Flood**, **Hurricane/Typhoon**, **Tornado**, etc.
- But, some other events such as **Drought**, **River Flood** and **Ice Storm** are introduced in the list of top 10 events with respect to ***Total Damage*** due to their high **Crop Damage** contribution. Whereas, other events such as **Tropical Storm**, **Winter Storm** and **High Wind** are eliminated.

Conclusion

This analysis gives an overview of the types of events which are most harmful with respect to population health and the types of events which have the greatest economic consequences.

Different types of events have different effects, for example **Excessive Heat** causes high number of fatalities while **Storm Surge** results in great economic consequences.

But, there are also some key events, such as **Tornado** and **Flood** which are both greatly harmful to public health as well as have high economic consequence. And, hence, these are recommonded to be the top priority during preparation for disaster management and mitigation.

End