# Extreme weather events in U.S. and their health and economic consequences

## Synopsis

In this report, we investigate extreme weather events and their consequences on public health and damage expenses. We use the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database to answer the two following questions

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

This document is divided as follows. The first section explains how we processed the data in a fully reproducible manner. In a second section we exposed our analysis and give the results. Finally we draw the conclusions of our analysis.

## Data collection and processing

The data are collected from the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database which can be downloaded at

https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2

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. It contains many variables related to extreme events. In view of our analysis, only a few of them are needed:

| Variables | Content |
|---|---|
| EVTYPE | Name of the event |
| BGN_DATE | Starting date of the event |
| FATALITIES | Number of fatalities caused by this weather event |
| INJURIES | Number of injuries |
| PROPDMG | Property damage amounts without units (in dollars) |
| PROPDMGEXP | Units as power of 10 for PROPDMG. Either an exponent (0,1,2,3,4,5,6,7,8) or a character: hecto (h or H = 10^2), kilo (K=10^3), mega (m or M = 10^6) and billion (B = 10^9). Other characters "", "-", "?", "+" will be set to 0. |
| CROPDMG | Crop damage amounts without units (in dollars) |
| CROPDMGEX | Units as power of 10 for for CROPDMG. Either an exponent (0,2) or a character: kilo (k, K=10^3), mega (m or M = 10^6) and billion (B = 10^9). Other characters "", "?" will be set to 0. |

We subset the initial dataset into a data frame named **data** and then **dataEVTYPE**. The variable **BGN_DATE** is useful if one wants to determine the impact of each events over different period of time.

```
library(rvest)
```

```
## Loading required package: xml2
```

```
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(ggplot2)


setwd("~/[coursera]DataSciences/DataScienceSpecialisation/5_Reproducible_research/RepData_PeerAssessmen

fileURL <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
fileName <- "StormData.csv.bz2"

if(!file.exists(fileName)){
        print("download")
        download.file(fileURL,  destfile = fileName, method="curl")
}

if(!(exists("dataset", envir = environment()))){
        print("dataset does not exist")
        dataset <- tbl_df(read.csv(fileName, header = TRUE, sep=','))
}
```

```
## [1] "dataset does not exist"
```

```
data <- dataset %>%
        select(EVTYPE, BGN_DATE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP)
```

Then, we convert **PROPDMGEXP** and **CROPDMGEXP** into their corresponding numerical values. The total impact of extreme weather events on public health is computed summing the number of fatalities and injuries. Similarly, the economic impact is found by summing the property and crop damages cost (converted into the right units). The rows with zero values for both entries are discarded as they will not influence the final result.

```
facMul<- c(1,0,0,0,1,10, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^2, 10^2, 10^3, 10^6, 10^6)
levels(data$PROPDMGEXP) <- facMul
facMulCrop<- c(1,0,1, 10^2, 10^9, 10^3, 10^3, 10^6, 10^6)
levels(data$CROPDMGEXP) <- facMulCrop

data <- data%>%
        mutate(POPHEALTH = FATALITIES + INJURIES)%>%
        mutate(DMEXPENSE = PROPDMG*as.numeric(levels(PROPDMGEXP))[PROPDMGEXP] + CROPDMG*as.numeric(leve
        filter(POPHEALTH >0 | DMEXPENSE >0)
```

The variable **EVTYPE** is particularly important as it determines the type of event. However, the same type of event can be stored under different names due to misspelling, use of synonym etc. We tidy up the data for the most important events regarding our two questions. Then for example, TSTM, THUNDERSTROM, SEVERE THUNDERSTORM ... all refer to THUNDERSTORM.

Another important modification of our database is related to TYPHOON/HURRICANE/CYCLONES. According to the NOAA website (http://oceanservice.noaa.gov/facts/cyclone.html): “Hurricanes, cyclones, and typhoons are all the same weather phenomenon; we just use different names for these storms in different places. In the Atlantic and Northeast Pacific, the term”hurricane" is used. The same type of disturbance in the Northwest Pacific is called a “typhoon” and “cyclones” occur in the South Pacific and Indian Ocean."

Therefore we consider them to be the same event.

```
dataEVTYPE <- data%>%
        mutate(EVTYPE = toupper(EVTYPE))%>%
        mutate(EVTYPE = gsub("[[:space:][:punct:]+]", " ", EVTYPE))%>%
        mutate(EVTYPE = gsub("TSTM|TH*UND*ER*[A-Z]*RMW*|THUNDERSTROM|THUDERSTORM",
                     "THUNDERSTORM", EVTYPE))%>%
        mutate(EVTYPE =gsub("THUNDERSTORM.*|SEVERE THUNDERSTORM", "THUNDERSTORM", EVTYPE))%>%
        mutate(EVTYPE =gsub("TORN.*", "TORNADO", EVTYPE))%>%
        mutate(EVTYPE =gsub("^BLIZZARD.*", "BLIZZARD", EVTYPE))%>%
        mutate(EVTYPE =gsub("EXCESSIVE|EXCESSIVELY|EXTREMELY|RECORD", "EXTREME",EVTYPE))%>%
        mutate(EVTYPE =gsub("DROUGHT|^EXTREME.*HEAT.*|^HEAT.*", "EXTREME HEAT", EVTYPE))%>%
        mutate(EVTYPE =gsub(".*FLOOD.*", "FLOODING", EVTYPE))%>%
        mutate(EVTYPE =gsub("ICE STORM.*|WINTER STORM.*", "ICE/WINTER STORM", EVTYPE))%>%
        mutate(EVTYPE =gsub("WILD.*FIRE.*", "WILD FIRE", EVTYPE))%>%
        mutate(EVTYPE =gsub("CURRENTS", "CURRENT", EVTYPE))%>%
        mutate(EVTYPE =gsub(".*WIND.*CHILL.*|^COLD", "EXTREME COLD", EVTYPE))%>%
        mutate(EVTYPE =gsub("^HIGH WINDS|STRONG.*WIND|^WIND", "HIGH WIND", EVTYPE))%>%
        mutate(EVTYPE =gsub("^HURRICANE.*| ^TYPH.*", "HURRICANE", EVTYPE))%>%
        mutate(EVTYPE =gsub("DENSE FOG", "FOG", EVTYPE))%>%
        mutate(EVTYPE =gsub(".*HIGH.*WIND.*", "HIGH WIND", EVTYPE))%>%
        mutate(EVTYPE =gsub("^LIGHTN.*", "LIGHTNING", EVTYPE))%>%
        mutate(EVTYPE =gsub("^SNOW.*", "HEAVY SNOW", EVTYPE))


glimpse(dataEVTYPE)
```

```
## Observations: 254,628
## Variables: 10
## $ EVTYPE     <chr> "TORNADO", "TORNADO", "TORNADO", "TORNADO", "TORNAD...
## $ BGN_DATE   <fctr> 4/18/1950 0:00:00, 4/18/1950 0:00:00, 2/20/1951 0:...
## $ FATALITIES <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 4, 0, ...
## $ INJURIES   <dbl> 15, 0, 2, 2, 2, 6, 1, 0, 14, 0, 3, 3, 26, 12, 6, 50...
## $ PROPDMG    <dbl> 25.0, 2.5, 25.0, 2.5, 2.5, 2.5, 2.5, 2.5, 25.0, 25....
## $ PROPDMGEXP <fctr> 1000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, 10...
## $ CROPDMG    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ CROPDMGEXP <fctr> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
## $ POPHEALTH  <dbl> 15, 0, 2, 2, 2, 6, 1, 0, 15, 0, 3, 3, 27, 12, 6, 54...
## $ DMEXPENSE  <dbl> 25000, 2500, 25000, 2500, 2500, 2500, 2500, 2500, 2...
```

## Results

In the previous section, we processed the data and we now use them to answer our two important questions.

### Most harmful weather events on injuries and fatalities

We create here a new dataframe **dataPopHealth** containing three variables: the type of event **EVTYPE**, the sum of all injuries and fatalities **TOTALPOPHEALTH** for a given event and the time period **PERIOD**. Then we display a barplot for the ten most catastrophic events for population health and it is clearly seen on the figure that TORNADO is the most harmful event from 1950 to present.

However according to the U.S. National Oceanic and Atmospheric Administration's website, only tornado events were recorded from 1950 through 1954. From 1955 through 1992, only tornado, thunderstorm wind and hail events were recorded as well.

Then it is interesting to perform the same analysis on different time period to verify that TORNADO is actually the most harmful weather event for population. We do the same analysis over the period from 1997-present. It confirms that TORNADOES are still the most harmfull event for population health ahead of extreme heat. However the proportion is significantly lower.

```
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:ggplot2':
##
##     ggsave
```

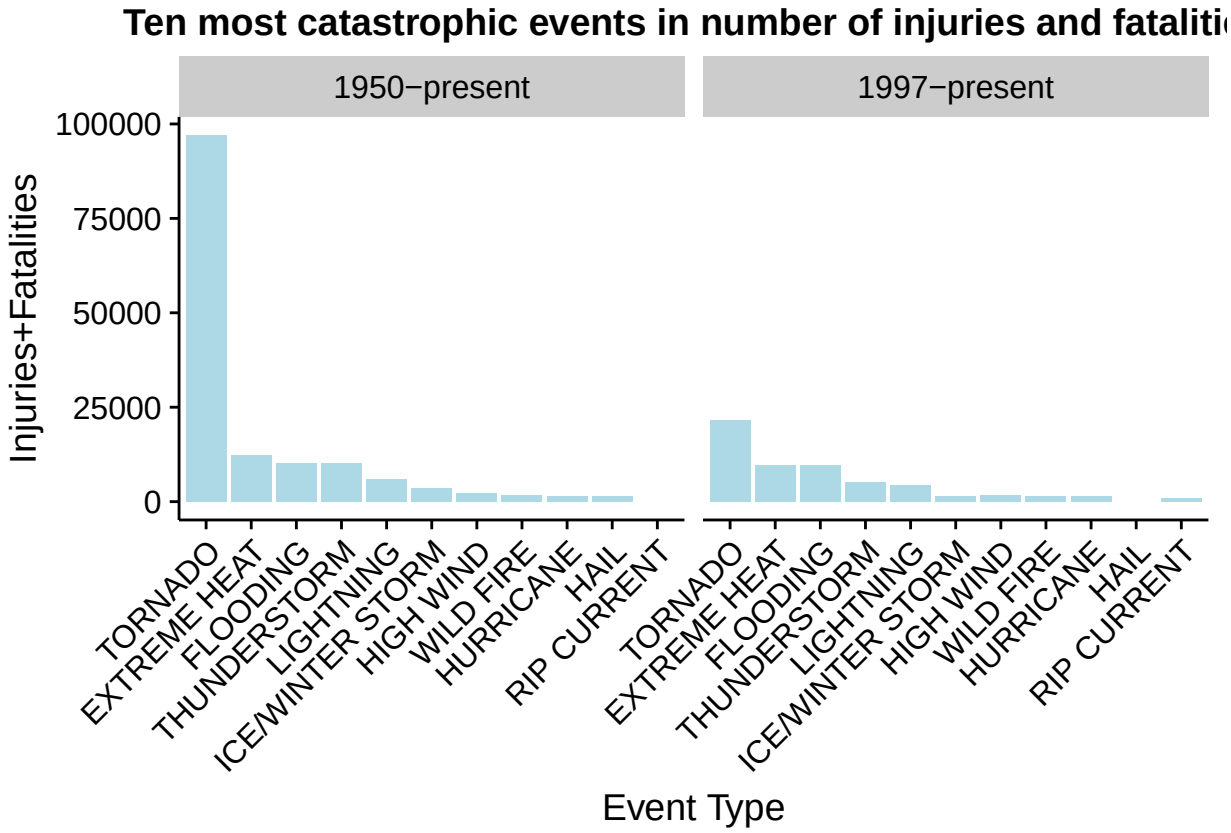
```
dataPopHealth <- dataEVTYPE %>%
        group_by(EVTYPE)%>%
        summarise(TOTALPOPHEALTH = sum(POPHEALTH, na.rm=TRUE))%>%
        arrange(desc(TOTALPOPHEALTH))%>%
        slice(1:10)%>%
        mutate(PERIOD="1950-present")


dataPopHealthRecent <- dataEVTYPE %>%
        mutate(BGN_DATE = strptime(BGN_DATE, format = "%m/%d/%Y %H:%M:%S") %>% as.character())%>%
        filter(BGN_DATE >= "1997-01-01")%>%
        group_by(EVTYPE)%>%
        summarise(TOTALPOPHEALTH = sum(POPHEALTH, na.rm=TRUE))%>%
        arrange(desc(TOTALPOPHEALTH))%>%
        slice(1:10)%>%
        mutate(PERIOD="1997-present")


combineALL <- bind_rows(dataPopHealth, dataPopHealthRecent)%>%
                mutate(EVTYPE = reorder(EVTYPE, -TOTALPOPHEALTH))

plot1 <- ggplot(combineALL, aes(x=EVTYPE,y=TOTALPOPHEALTH, group=PERIOD))+
        geom_bar(position = "dodge",stat="identity", fill="lightblue")+
        facet_grid(.~ PERIOD)+
        labs(x="Event Type", y="Injuries+Fatalities") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        labs(title="Ten most catastrophic events in number of injuries and fatalities")


print(plot1)
```

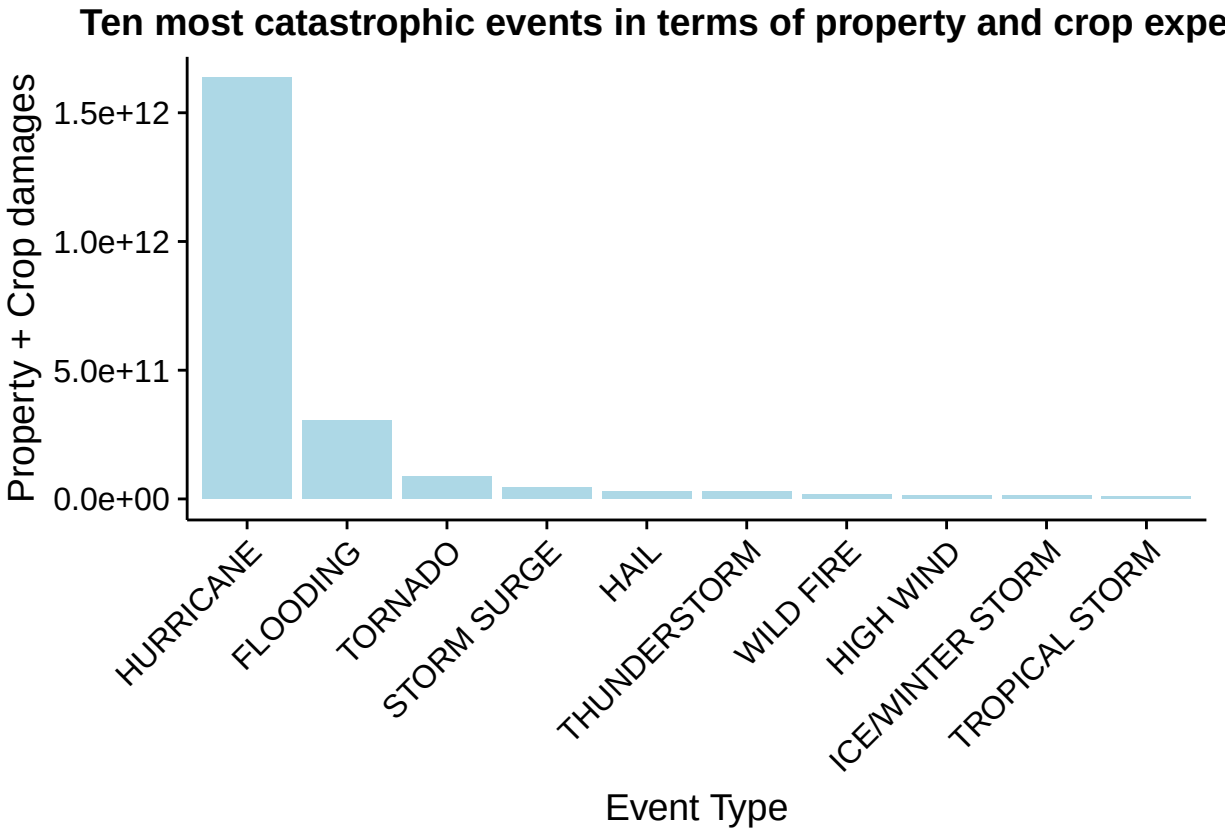



**Most harmul weather events on property and crop damages**

We now investigate the economical impact of those events on properties and crops, stored in the variable DMEXPENSE. In this perspective, HURRICANE are the most harmful events. We do the analysis only for the full period (1997 to present) since HURRICANE have only been registered recently from 1997.

```
dataECO <- dataEVTYPE %>%
        group_by(EVTYPE)%>%
        summarise(TOTALDMEXPENSE = sum(DMEXPENSE, na.rm=TRUE))%>%
        arrange(desc(TOTALDMEXPENSE))%>%
        mutate(EVTYPE = reorder(EVTYPE, -TOTALDMEXPENSE))


plot2 <- ggplot(slice(dataECO, 1:10), aes(x=EVTYPE,y=TOTALDMEXPENSE))+
        geom_bar(position = "dodge",stat="identity", fill="lightblue")+
        labs(x="Event Type", y="Property + Crop damages") +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
        labs(title="Ten most catastrophic events in terms of property and crop expenses")
print(plot2)
```




## Conclusion

Our analysis shows that TORNADO are more harmfull to people and HURRICANE to properties. It would be interesting to understand why TORNADOES are more dangerous to people. A reason might be that they occur near big cities more often than HURRICANE does for example.