Analysis of severe weather events in the united states between the years 1950 and 2011

MATIAS, V. A.

28/03/2021

1. Synopsis

This is the final project of the Reproducible Research Course offered by Johns Hopkins University on the Coursera platform.

This paper will present an analysis of climate events based on the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database.

The programming language R will be used as the basis for all the analysis to be presented, this document being developed with R markdown (Rmd).

1.1. Data

The available data are in a compressed CSV file for type .bz2, resulting in a 42Mb file.

Data: Storm Data

Documentation: National Weather Service Storm Data Documentation

FAQ: National Climatic Data Center Storm Events FAQ

2. Data Processing

This section covers the techniques used in the collection and processing of data

2.1. Required libraries

Before starting the analysis, some packages will be needed to facilitate development:

- dplyr: To use the function mapvalues, a good way to use the idea of hashmap in R
- dplyr: To facilitate data manipulation
- ggplot2: To create the graphics
- gridExtra: To create a view with more than one panel
- data.table: To use the *fread()* function, being faster to read a CSV than the *read.csv()* function
- R.utils: To allow the fread function to read a CSV without having to unzip a bz2 file

```
library(plyr) # Must be called before dplyr
library(dplyr)
library(ggplot2)
library(gridExtra)
library(data.table)
library(R.utils)
```

2.2. Geting the data

The project requires that the compressed CSV files be in the directory where the analysis will be performed. This step can be performed by the user himself, downloading the file and placing it in the analysis directory, or by the code below.

```
# Optional step
data_url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
download.file(data_url, destfile = "storm_data.csv.bz2")
```

2.3. Loading the data

Now the downloaded file with the name *storm_data.csv.bz2* will be stored in memory. Note that this is a large file when uncompressed, requiring almost 500Mb to allocate it in RAM.

To avoid storing 500Mb in RAM, we will only select the columns that will be used in the analysis. They are: PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP, INJURIES, FATALITIES, EVTYPE. This approach results in just under 50Mb to be stored in memory.

```
weather <- fread("storm_data.csv.bz2") %>%
  as_tibble() %>%
  select(PROPDMG,
         PROPDMGEXP,
         CROPDMG,
         CROPDMGEXP,
         INJURIES,
         FATALITIES,
         EVTYPE) %>%
  mutate(EVTYPE = as.factor(EVTYPE))
object.size(weather)
```

```
## 47004032 bytes
```

3. Results

The analysis of the data and its consequent results comes from the questions proposed by the Assignment.

3.1. Harmful events

The first analysis to be made is: given the different types of recorded climatic events, which are the most harmful to the population health? The code below realized some transformations in the original data to get the events with more frequencies of deaths and injuries. At the end is created one figure composed of two histograms.

```r
#Event types with most injured people
injuries_evtype <- weather %>%
  group_by(EVTYPE) %>%
  summarise(sum_injuries = sum(INJURIES), .groups = "drop") %>%
  arrange(desc(sum_injuries)) %>%
  head()
injuries_evtype
```

```
## # A tibble: 6 x 2
##   EVTYPE         sum_injuries
##   <fct>                 <dbl>
## 1 TORNADO               91346
## 2 TSTM WIND              6957
## 3 FLOOD                  6789
## 4 EXCESSIVE HEAT         6525
## 5 LIGHTNING              5230
## 6 HEAT                   2100
```

```r
#Event types with most fatalities
fatal_evtype <- weather %>%
  group_by(EVTYPE) %>%
  summarise(sum_fatalities = sum(FATALITIES), .groups = "drop") %>%
  arrange(desc(sum_fatalities)) %>%
  head()
fatal_evtype
```

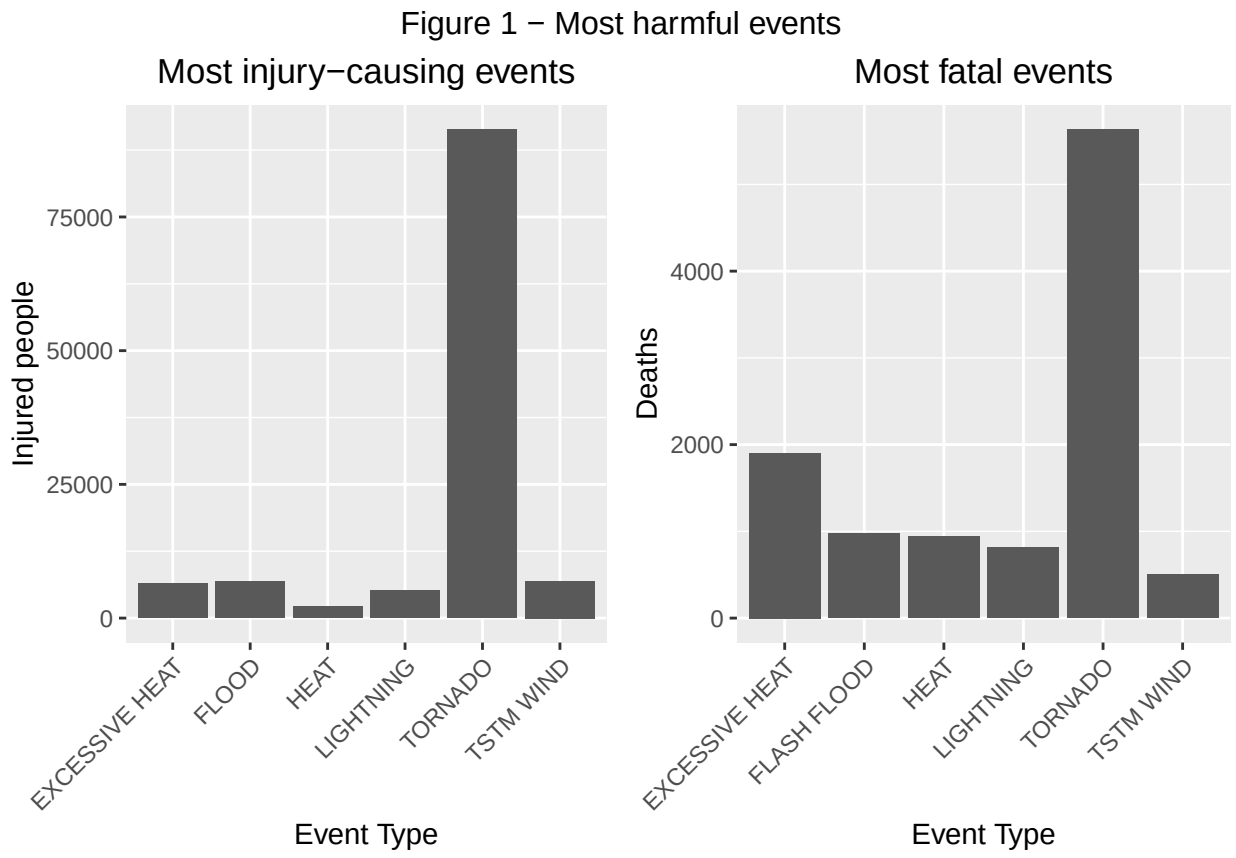
```
## # A tibble: 6 x 2
##   EVTYPE         sum_fatalities
##   <fct>                   <dbl>
## 1 TORNADO                  5633
## 2 EXCESSIVE HEAT           1903
## 3 FLASH FLOOD               978
## 4 HEAT                      937
## 5 LIGHTNING                 816
## 6 TSTM WIND                 504
```

```r
#Bar plot of the events with the most injuries
injuries_plot <- ggplot(injuries_evtype, aes(x = EVTYPE, y = sum_injuries)) +
  geom_bar(stat = "identity") +
  labs(x = "Event Type",
       y = "Injured people",
       title = "Most injury-causing events") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))

#Bar plot of the events with the most deaths
fatal_plot <- ggplot(fatal_evtype, aes(x = EVTYPE, y = sum_fatalities)) +
  geom_bar(stat = "identity") +
  labs(x = "Event Type",
       y = "Deaths",
       title = "Most fatal events") +
  theme(plot.title = element_text(hjust = 0.5),
```

```
        axis.text.x = element_text(angle = 45, hjust = 1))

#Figure of the most harmful events
gridExtra::grid.arrange(injuries_plot, fatal_plot,
                        nrow = 1,
                        top = "Figure 1 - Most harmful events")
```



As we can see, the six events with more causes of injuries also are the same events with more causes of deaths. Tornadoes are the most harmful events pointed out by the data.

3.2. Economic consequences

The second question is to analyze which climatic events have caused the most economic damage.

From the review of the database documentation and the analysis of the variables in the CSV file, it is possible to identify that the PROPDMG attribute refers to the damage caused on properties, just as CROPDMG refers to the damage caused on crops.

The documentation refers to factors K, M and B as well. These factors mean thousands, millions and billions, respectively. The PROPDMGEXP and CROPDMGEXP variables store these references for the loss values. There are still other values not specified in the documentation but present in these columns (such as numerical values and letters with no relation, such as “H”), these values were disregarded because there was no way to measure them.

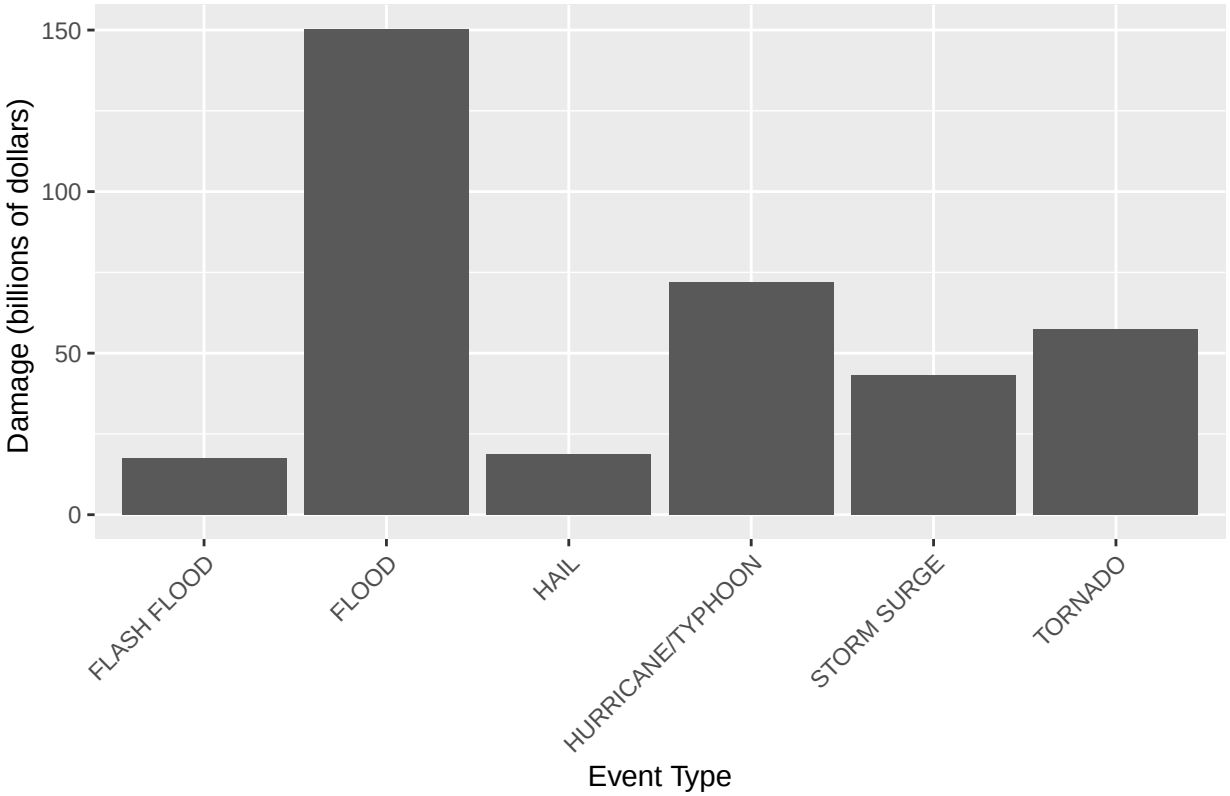
```r
# The two vectors below will be used in the function mapvalues, working like
# a numeric reference for each character (K = 10^3, M = 10^6 and B = 10^9)
units_known <- c("K", "M", "B")
ref_known <- c(10^3, 10^6, 10^9)

#Creating the data with the value in dollars (2011) of damage in crops and properties
dmg <- weather %>%
  filter(PROPDMGEXP %in% units_known | CROPDMGEXP %in% units_known) %>%
  mutate(PROPDMGEXP = mapvalues(PROPDMGEXP, from = units_known, to = ref_known),
         CROPDMGEXP = mapvalues(CROPDMGEXP, from = units_known, to = ref_known),
         PROPDMG = PROPDMG * as.numeric(PROPDMGEXP),
         CROPDMG = CROPDMG * as.numeric(CROPDMGEXP)) %>%
  group_by(EVTYPE) %>%
  summarise(sum_propdmg = sum(PROPDMG, na.rm = TRUE),
            sum_cropdmg = sum(CROPDMG, na.rm = TRUE),
            total_dmg = sum_propdmg + sum_cropdmg,
            .groups = "drop") %>%
  arrange(desc(total_dmg))
dmg
```

```
## # A tibble: 429 x 4
##    EVTYPE             sum_propdmg sum_cropdmg    total_dmg
##    <fct>                    <dbl>       <dbl>        <dbl>
##  1 FLOOD             144657709800  5661968450 150319678250
##  2 HURRICANE/TYPHOON  69305840000  2607872800  71913712800
##  3 TORNADO            56925660480   414953110  57340613590
##  4 STORM SURGE        43323536000        5000  43323541000
##  5 HAIL               15727366870  3025537450  18752904320
##  6 FLASH FLOOD        16140811542  1421317100  17562128642
##  7 DROUGHT             1046106000 13972566000  15018672000
##  8 HURRICANE          11868319010  2741910000  14610229010
##  9 RIVER FLOOD         5118945500  5029459000  10148404500
## 10 ICE STORM           3944927810  5022113500   8967041310
## # ... with 419 more rows
```

```r
#Making the plot (here we will use just the 6 events with most damage)
ggplot(head(dmg), aes(x = EVTYPE, y = total_dmg/10^9)) +
  geom_bar(stat = "identity") +
  labs(x = "Event Type",
       y = "Damage (billions of dollars) ",
       title = "Figure 2 -Types of events with the greatest economic consequences") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 2 −Types of events with the greatest economic consequences



Although tornadoes have the greatest responsibility for the damage to the health of the population, they account for just over one third of the greatest cause of economic damage: Floods. It is worth noting that, even so, tornadoes are the third biggest cause of economic losses.