Peer Review Assesment 2

Ali Unlu

6/10/2021

1 - Synopsis

In this report, we will analysis the data storm data provided by the Johns Hopkins University

My data analysis aims to address the following questions:

1. Across the United States, which types of events (as indicated in the **EVTYPE** variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

2 - Data Processing

Data preperation

From a list of variables in storm.data, these are columns of interest:

Health variables:

FATALITIES: approx. number of deaths
INJURIES: approx. number of injuries

Economic variables:

PROPDMG: approx. property damags
PROPDMGEXP: the units for property damage value
CROPDMG: approx. crop damages
CROPDMGEXP: the units for crop damage value

Events - target variable:

EVTYPE: weather event (Tornados, Wind, Snow, Flood, etc..)

```
# we load the data from computer
raw.data <- read.csv(bzfile("StormData.csv.bz2"))

# required packages for this analysis
library(plyr)
library(dplyr)
library(ggplot2)
library(reshape2)
```

We kept only relevant variables for this analysis

```
# check out column names
names(raw.data)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

```
# subset EVTYPE and cost related variables
raw.data <- raw.data[c(8, 23:28)]


# Only cases with fatalities or injuries occurred.
raw.data <- subset(raw.data, EVTYPE != "?" &  INJURIES > 0 | FATALITIES > 0 | PROPDMG > 0 | CROPDMG > 0)
```

3 - Converting the exponent columns (PROPDMGEXP and CROPDMGEXP)

```
# first, we will check the properties
table(raw.data$PROPDMGEXP)
```

```
##
##             -      +      0      2      3      4      5      6      7      B
##  11585      1      5    210      1      1      4     18      3      3     40
##      h      H      K      m      M
##      1      6 231428      7  11320
```

```
table(raw.data$CROPDMGEXP)
```

```
##
##             ?      0      B      k      K      m      M
## 152664      6     17      7     21  99932      1   1985
```

```
# some variables have lower case values
# we are now converting lower cases to uppercases in these variables
raw.data <-  data.frame(lapply(raw.data, function(v) {
        if (is.character(v)) return(toupper(v))
        else return(v)
        }))
```

According to the previous tables, the CROPDMGEXP only contains a subset of these values. Most of the numerical exponents are missing. The factor is only calculated for the exponents provided in that variable.

There is some mess in units, so we transform those variables in one unit (dollar) variable by the following rule:
+ K or k: thousand dollars (10^3)
+ M or m: million dollars (10^6)
+ B or b: billion dollars (10^9)
+ the rest would be consider as dollars

```
#  for PROPDNGEXP
raw.data2 <- raw.data
raw.data2$PROPDMGEXP <- mapvalues(raw.data2$PROPDMGEXP,
        from = c("K","M","", "B", "m", "+", "0","5", "6", "?", "4", "2", "3", "h", "7","H", "-","1", "8
        to = c(10^3,10^6,10^0,10^9,10^6,10^0,10^0,10^5,10^6,10^0,10^4,10^2,10^3,10^2,10^7,10^2,10^0,10^

raw.data2$PROPDMGEXP <- as.numeric(as.character(raw.data2$PROPDMGEXP))
# total PROPDMGEXP
raw.data2$PROPDMGTOTAL <- (raw.data2$PROPDMG * raw.data2$PROPDMGEXP)/1000000000

# for CROPDMGEXP
raw.data2$CROPDMGEXP <- mapvalues(raw.data2$CROPDMGEXP,
                                  from = c("",   "M",  "K",   "m",  "B",  "?",    "0",     "k",   "2"),
                                  to =   c(10^0, 10^6, 10^3,  10^6, 10^9,  10^0,   10^0,   10^3,  10^2)
raw.data2$CROPDMGEXP <- as.numeric(as.character(raw.data2$CROPDMGEXP))

# total CROPDMGEXP
raw.data2$CROPDMGTOTAL <- (raw.data2$CROPDMG * raw.data2$CROPDMGEXP)/1000000000
```

4 Results

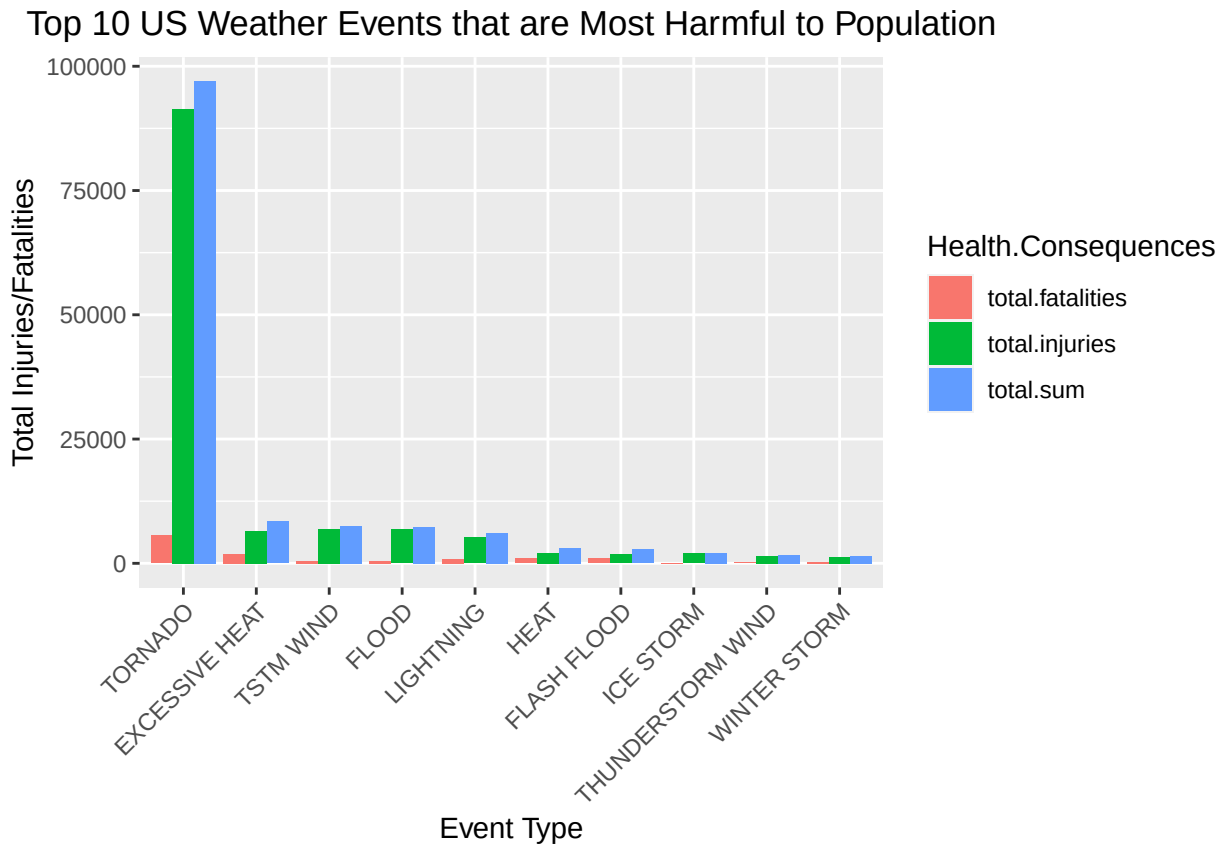
Calculating Total Fatalities and Injuries

Table of public health problems by event type

```
# We sum the fatalities and injuries to calculate the total effects of events on public health
mergedFI <- raw.data2 %>%
        group_by(EVTYPE)%>%
        summarise(total.fatalities=sum(FATALITIES), total.injuries= sum(INJURIES), total.sum= total.fata
        arrange(desc(total.sum))

mergedFI10 <- mergedFI[1:10,]

# check the list
head(mergedFI10, 5)
```

```
## # A tibble: 5 x 4
##   EVTYPE         total.fatalities total.injuries total.sum
##   <chr>                     <dbl>          <dbl>     <dbl>
## 1 TORNADO                    5633          91346     96979
## 2 EXCESSIVE HEAT             1903           6525      8428
## 3 TSTM WIND                   504           6957      7461
## 4 FLOOD                       470           6789      7259
## 5 LIGHTNING                   816           5230      6046
```

```
############## plot
#gplot
# Melting data so that it is easier to put in bar graph format
Health_Consequences <- melt(mergedFI10, id.vars = "EVTYPE", variable.name = "Health.Consequences")

ggplot(Health_Consequences, aes(x = reorder(EVTYPE, -value), y = value)) +
        geom_bar(stat = "identity", aes(fill = Health.Consequences), position = "dodge") +
        ylab("Total Injuries/Fatalities") +
        xlab("Event Type") +
        theme(axis.text.x = element_text(angle=45, hjust=1)) +
        ggtitle("Top 10 US Weather Events that are Most Harmful to Population") +
        theme(plot.title = element_text(hjust = 0.5))
```

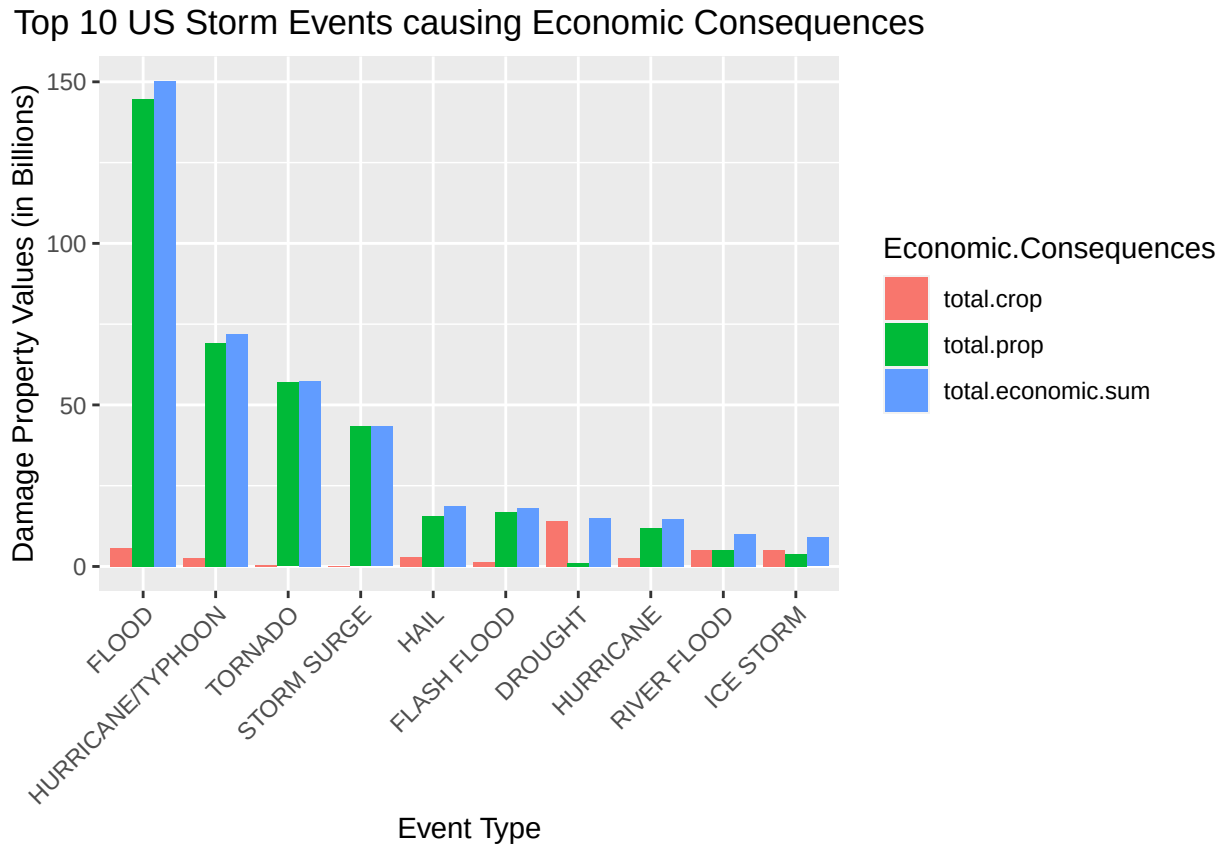


The barchart shows that ***Tornados*** are the most harmful weather events for people's health.

Estimating the total of Property Cost and Crop Cost (Economic Impacts)

```
############## Economic costs
# We sum the fatalities and injuries to calculate the total effects of events on public health
total.economic <- raw.data2 %>%
        group_by(EVTYPE)%>%
        summarise(total.crop= sum(CROPDMGTOTAL), total.prop= sum(PROPDMGTOTAL), total.economic.sum= tota
        arrange(desc(total.economic.sum))
```

```r
economic10 <- total.economic[1:10,]


# plot
# Melting data so that it is easier to put in bar graph format
Economic.Consequences <- melt(economic10, id.vars = "EVTYPE", variable.name = "Economic.Consequences")
ggplot(Economic.Consequences, aes(x = reorder(EVTYPE, -value), y = value)) +
        geom_bar(stat = "identity", aes(fill = Economic.Consequences), position = "dodge") +
        ylab("Damage Property Values (in Billions)") +
        xlab("Event Type") +
        theme(axis.text.x = element_text(angle=45, hjust=1)) +
        ggtitle("Top 10 US Storm Events causing Economic Consequences") +
        theme(plot.title = element_text(hjust = 0.5))
```



The barchart shows that ***Floods*** cause the biggest economical damages.