Analysis for the influences of different weather conditions

Haralampi Bageski

Sunday, October 26, 2014

```
setwd("C:/Users/hari/Documents/R/Reproducible Research/Assignment2")
library(utils)
#downloading...
download.file(url="https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",dest="repda
```

```
## Warning: running command 'curl
## "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
## -o "repdata-data-StormData.csv.bz2"' had status 127
```

```
## Warning in download.file(url =
## "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",
## : download had nonzero exit status
```

```
#unzipping and reading the data...
Storm_Data <- read.csv("repdata-data-StormData.csv.zip",stringsAsFactors = FALSE)
```

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

First we are going to group the weather types and count the total fatalities and injuries. At the same time we will sort them first by total fatalities (as a more important factor in population health), and second by total injuries in decreasing order.

```
library(sqldf)
```

```
## Loading required package: gsubfn
## Loading required package: proto
## Loading required package: RSQLite
## Loading required package: DBI
## Loading required package: RSQLite.extfuns
```

```
#figure out the names of the columns:
names(Storm_Data)
```

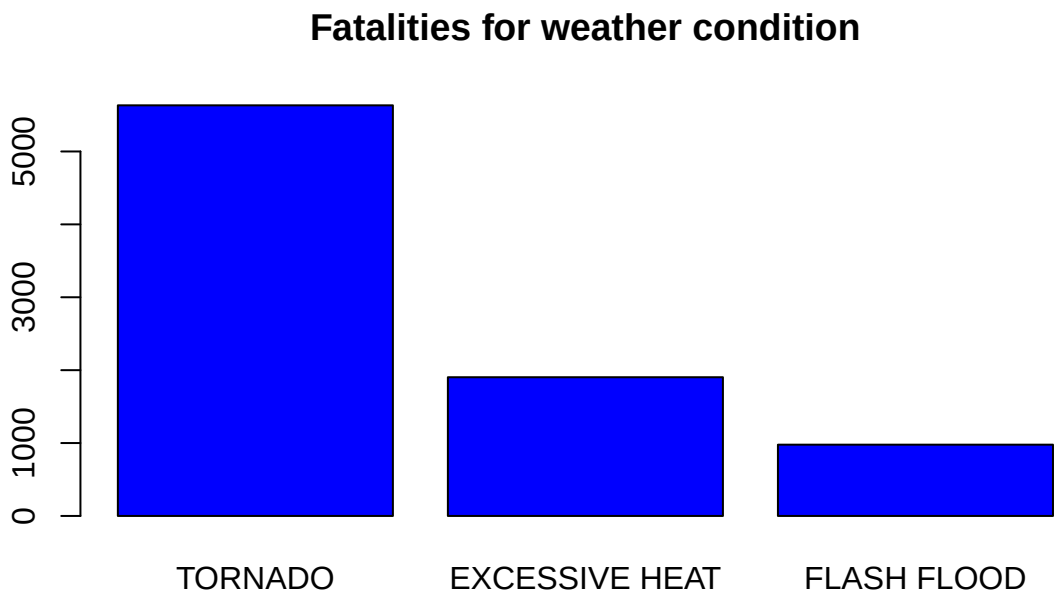
```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"     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

```
#deleting the columns that are not needed
Storm_Data <- Storm_Data[,c(-(1:7),-(9:22),-(29:37))]
#do the grouping by weather condition to get the total fatailities and injuries:
Health_damage <- sqldf("select EVTYPE,sum(FATALITIES) as TotalFatalities,sum(INJURIES) as TotalInjuries
```

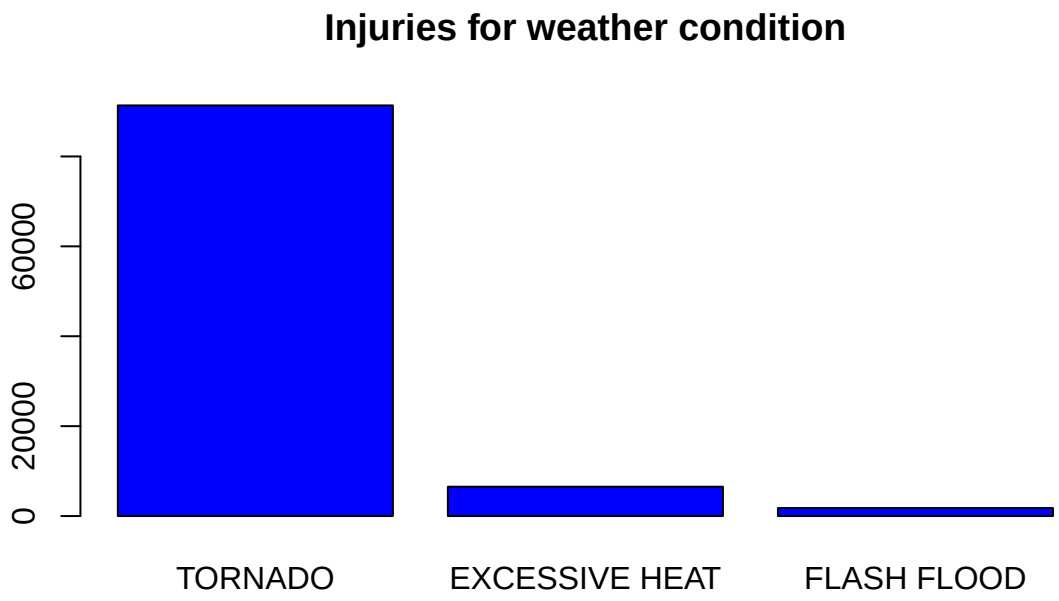
```
## Loading required package: tcltk
```

Two figures follow that represent the results for the 3 most severe weather conditions. The first one presents the number of fatalities for each weather condition. The second presents the number of injuries for each weather condition. For simplicity I regard only the most terrible conditions.

```
barplot(Health_damage$TotalFatalities[1:3],names.arg=Health_damage$EVTYPE[1:3], col = "blue",main= "Fata
```




```
barplot(Health_damage$TotalInjuries[1:3],names.arg=Health_damage$EVTYPE[1:3], col = "blue",main = "Inju
```



2. Across the United States, which types of events have the greatest economic consequences?

We need to do some data processing for the columns "PROPDMG","PROPDMGEXP","CROPDMG","CROPDMGEXP" which tell us about the approximated damage done to properties and crops. We calculate the damage as DMG*10^DMGEXP. Since for this part of the assignment we should only care about the above mentioned four columns plus the one that shows the type of weather, we extract only those five. Then we must be careful about the empty fields.

```
EcoData <- sqldf("select EVTYPE,PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGEXP
                 from  Storm_Data")
unique(EcoData$PROPDMGEXP)
```

```
##  [1] "K" "M" ""  "B" "m" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "H" "-"
## [18] "1" "8"
```

Then we must normalize the DMGEXP columns since as we can see they contain some non-numeric values. We know that H/h=2 (in the above formula it would produce 10^2= a hundred), K/k=3, M/m=6, B/b=9. Other characters are '-' meaning at most,'+' means at least, '?' means unknown, and '' means empty data, so all of them can be substituted with '0' (we can use other strategies for the '?' but for simplicity we change it to '0').

```
#Change from letters for the degree to numbers :
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP ==""] <- 0
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="+"] <- 0
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="?"] <- 0
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="-"] <- 0
```

```
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="K"] <- 3
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="M"] <- 6
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =='B'] <- 9
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="m"] <- 6
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="h"] <- 2
EcoData$PROPDMGEXP[EcoData$PROPDMGEXP =="H"] <- 2
unique(EcoData$PROPDMGEXP)
```

```
##  [1] "3" "6" "0" "9" "5" "4" "2" "7" "1" "8"
```

```
unique(EcoData$CROPDMGEXP)
```

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

```
#Change from letters for the degree to numbers :
EcoData$CROPDMGEXP[EcoData$CROPDMGEXP ==""] <- 0
EcoData$CROPDMGEXP[EcoData$CROPDMGEXP =="?"] <- 0
EcoData$CROPDMGEXP[EcoData$CROPDMGEXP =="K"] <- 3
EcoData$CROPDMGEXP[EcoData$CROPDMGEXP =="M"] <- 6
EcoData$CROPDMGEXP[EcoData$CROPDMGEXP =="m"] <- 6
EcoData$CROPDMGEXP[EcoData$CROPDMGEXP =="B"] <- 9
EcoData$CROPDMGEXP[EcoData$CROPDMGEXP =="k"] <- 3
unique(EcoData$CROPDMGEXP)
```

```
## [1] "0" "6" "3" "9" "2"
```

Then we can create two more columns for the total cost of Properties and Crops:

```
options(scipen=999)
EcoData$PropCost <-as.numeric(as.character(EcoData$PROPDMG))*10^as.numeric(as.character(EcoData$PROPDMGE
EcoData$CropCost<-as.numeric(as.character(EcoData$CROPDMG))*10^as.numeric(as.character(EcoData$CROPDMGEX
```

Results

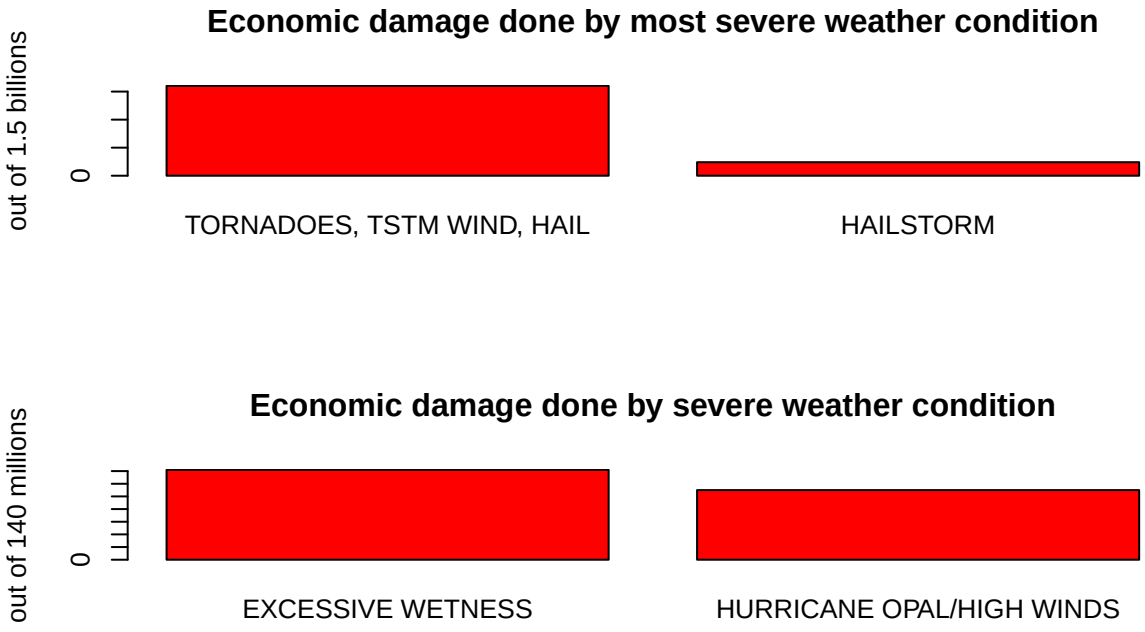
```
#do the grouping by in decreasingly order for the total cost:
Most_Damaging_to_Eco <- sqldf("select EVTYPE, sum(PropCost) as PropCostTotal,
                              sum(CropCost) as CropCostTotal,
                              sum(CropCost+PropCost) as TotalCost from EcoData
                              group by EVTYPE
                              order by PropCost+CropCost DESC")
#the most economically harmful is:
Most_Damaging_to_Eco$TotalCost[1] #1602500000
```

```
## [1] 1602500000
```

```
Most_Damaging_to_Eco$TotalCost[3] #142000000
```

```
## [1] 142000000
```

```r
#I insert two histogram of most damaging weather conditions
layout(matrix(c(1,1,2,2), 2, 2, byrow = TRUE))
barplot(Most_Damaging_to_Eco$TotalCost[1:2],
        names.arg=Most_Damaging_to_Eco$EVTYPE[1:2], col = "red",xpd=FALSE,
        ylab="out of 1.5 billions",
        main = "Economic damage done by most severe weather condition")
barplot(Most_Damaging_to_Eco$TotalCost[3:4],
        names.arg=Most_Damaging_to_Eco$EVTYPE[3:4], col = "red",xpd=FALSE,
        ylab="out of 140 millions",
        main = "Economic damage done by severe weather condition")
```



We can conclude that the tornadoes and excessive heat are most severe to the population health and hail combined with tornado or just itself is most severe towards the economy.