Data Science - Reproducible research - Course project 2

Analysis of the NOAA storm database to identify weather phenomena in the USA with the biggest impact on health and economy.

By Bert Lijnen July 2017

1. Introduction

The U.S. National Oceanic and Atmospheric Administration's (NOAA) storm 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. We will use this database in the analysis below to investigate which weather phenomena have the biggest impact on both heath and economy.

2. Synopsis

The analysis of the NOAA storm database revealed that between 1996 and 2011 excessive heat was the deedliest threat to the US population, followed by tornados. These results confirm that global warming needs to be taken seriously. In addition, of all weather events, tornados most often lead to injuries, followed at distance by flood. Both events also had the highest economic impact in the USA.

3. Data Processing

First we import the the raw CSV file containing the data. This document will be used for the data preprocessing

```
library(readr)
stormdata <- read_csv("~/repdata%2Fdata%2FStormData.csv")
```

We can reduce the dataset by eliminating records prior to 1996. As shown in the table below, the number of reports in earlier years is very low. The number significantly increased in the interval (1996-2001].

```
stormdata$YEAR <- as.numeric(format(as.Date(stormdata$BGN_DATE,
                                            format= "%m/%d/%Y%H:%M:%S"), "%Y"))
breaks<-cut(stormdata$YEAR,12)
table(breaks)
```

```
## breaks
## (1950,1955] (1955,1960] (1960,1965] (1965,1970] (1970,1975] (1975,1980]
##        3278        9858       11806       14529       20463       21578
## (1980,1986] (1986,1991] (1991,1996] (1996,2001] (2001,2006] (2006,2011]
##       35285       44706       87264      164838      189554      299138
```

We use the following code to extract the post 1996 data

```
stormdata<-stormdata[stormdata$YEAR>=1996,]
```

We can further reduce the dataset by retaining only the variables that are relevant for our analysis. To answer the research questions, we only need the following variables:

- EVTYPE: the type of storm and other signifcant weather phenomena (e.g. Dense fog, hurricane, lightning, tropical storm, etc.)

- FATALITIES:the number of people who died as a direct result of the storm
- INJURIES: the number of people who got injured as a direct result of the storm
- PROPDMG:the base amount of property damage
- PROPDMGEXP: the base multiplier for property damage.The letter values stand for H,h = hundreds K,k = thousands M,m = millions B,b = billions
- CROPDMG: the base amount of crop damage
- CROPDMGEXP: the base multiplier for crop damage.

```
needed_variables<-c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG",
                    "PROPDMGEXP" ,"CROPDMG", "CROPDMGEXP")
```

Now we apply both subsets to create the new dataset and check the result

```
stormnew<-as.data.frame(stormdata[needed_variables])
knitr::kable(head(stormnew))
```

EVTYPE	FATALITIES	INJURIES	PROPDMG	PROPDMGEXP	CROPDMG	CROPDMGEXP
WINTER STORM	0	0	380	K	38	K
TORNADO	0	0	100	K	0	NA
TSTM WIND	0	0	3	K	0	NA
TSTM WIND	0	0	5	K	0	NA
TSTM WIND	0	0	2	K	0	NA
HAIL	0	0	0	NA	0	NA

We now transform the exponential terms in variables PROPDMGEXP and CROPDMGEXP to actual values, but first we check what the unique values are of both variables to check what values we need to replace.

```
unique(stormnew$PROPDMGEXP)
```

```
## [1] "K" NA  "M" "B" "0"
```

```
unique(stormnew$CROPDMGEXP)
```

```
## [1] "K" NA  "M" "B"
```

```
stormnew[is.na(stormnew$PROPDMGEXP),]<-0
stormnew[is.na(stormnew$CROPDMGEXP),]<-0

stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="?"] <- 0
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="+"] <- 0
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="-"] <- 0
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP==""] <- 1
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="0"] <- 1
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="5"] <- 100000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="6"] <- 1000000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="4"] <- 10000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="2"] <- 100
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="3"] <- 1000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="8"] <- 100000000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="H"] <- 100
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="h"] <- 1
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="K"] <- 1000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="M"] <- 1000000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="m"] <- 1000000
stormnew$PROPDMGEXP[stormnew$PROPDMGEXP=="B"] <- 1000000000
```

```
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP=="?"] <- 0
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP==""] <- 1
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP=="2"] <- 100
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP=="K"] <- 1000
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP=="k"] <- 1000
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP=="M"] <- 1000000
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP=="m"] <- 1000000
stormnew$CROPDMGEXP[stormnew$CROPDMGEXP=="B"] <- 1000000000
```

3. Results

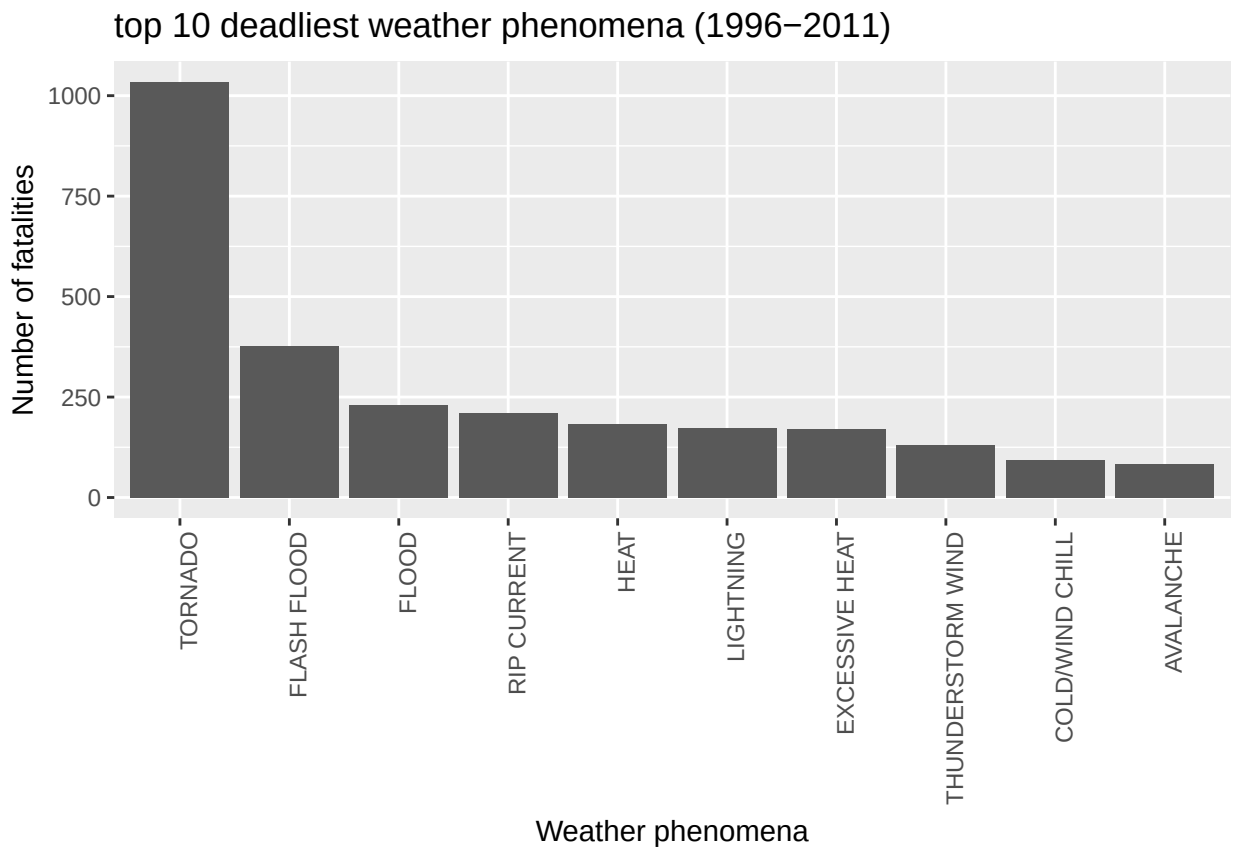
The first research question we address is which weather phenomena are most harmful with respect to population health ?

Two variables give information about the population health: the number of fatalities and the number of injuries. We will now compute which events caused most deaths and injuries. We are only interested in the events with a major impact, so we only select the 10 events causing most casualties and injuries.

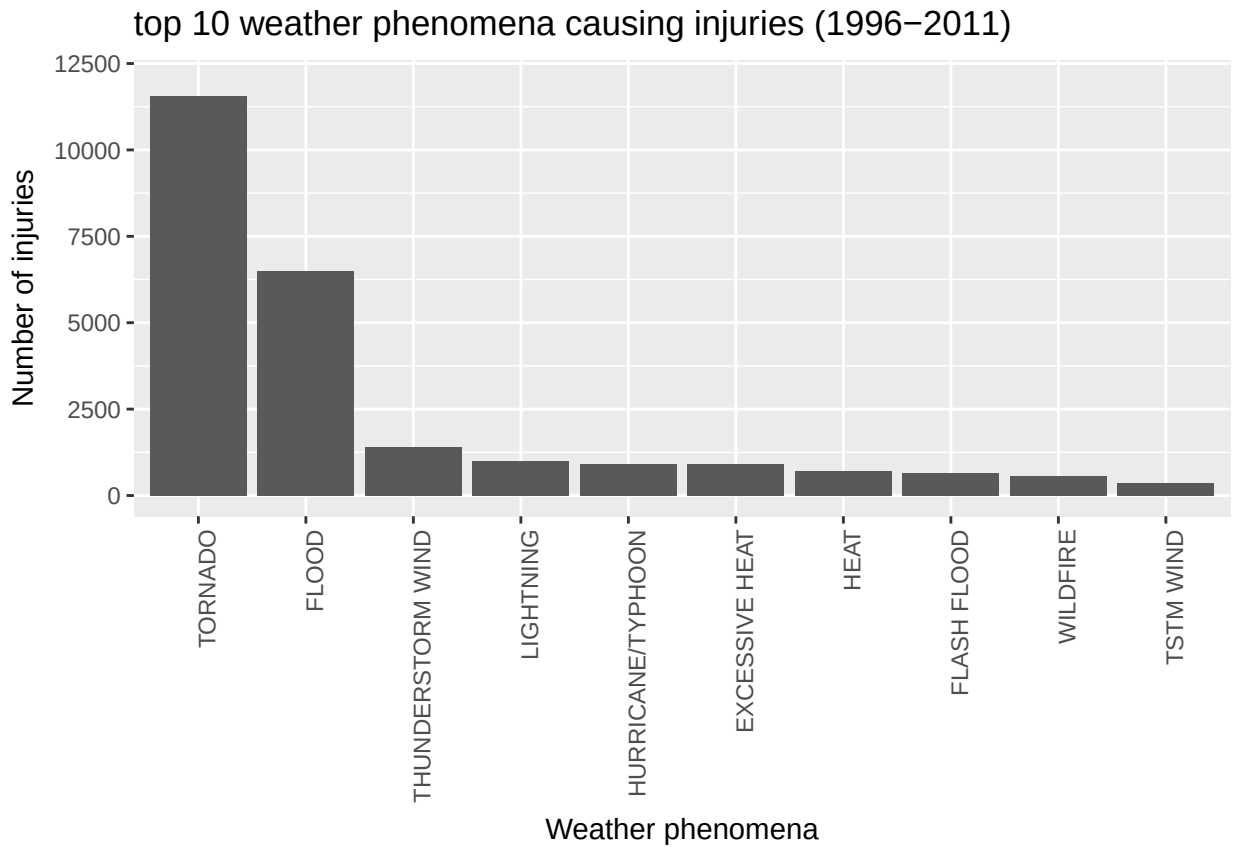
```
fatal <- aggregate(FATALITIES ~ EVTYPE, data = stormnew, sum)
injured <- aggregate(INJURIES ~ EVTYPE, data = stormnew, sum)
top10_fatal<-fatal[order(-fatal$FATALITIES),][1:10,]
top10_injured<-injured[order(-injured$INJURIES),][1:10,]
```

We will now visualise both top 10 rankings by means of a barplot.

```
library(ggplot2)
x<-ggplot()+geom_bar(data=top10_fatal, aes(x=reorder(top10_fatal$EVTYPE,-top10_fatal$FATALITIES),
                                           y=top10_fatal$FATALITIES),
                     stat="identity")
x<-x+theme(axis.text.x=element_text(angle=90, hjust=1))
x<-x+xlab("Weather phenomena")+ylab("Number of fatalities")+
        ggtitle("top 10 deadliest weather phenomena (1996-2011)")+
        theme(axis.text.x=element_text(angle=90, hjust=1))
print(x)
```




```
y<-ggplot()+geom_bar(data=top10_injured,
                     aes(x=reorder(top10_injured$EVTYPE,-top10_injured$INJURIES),
                         y=top10_injured$INJURIES), stat="identity")
y<-y+theme(axis.text.x=element_text(angle=90, hjust=1))+ylim(0,12000)
y<-y+xlab("Weather phenomena")+ylab("Number of injuries")+
        ggtitle("top 10 weather phenomena causing injuries (1996-2011)")+
        theme(axis.text.x=element_text(angle=90, hjust=1))
print(y)
```



The second research question we address is which weather phenomena have the greatest economic consequences. To answer this question we need the CROPDMG and PROPDMG data as well as their transformd exponential counterparts CROPDMGEXP and PROPDMGEXP

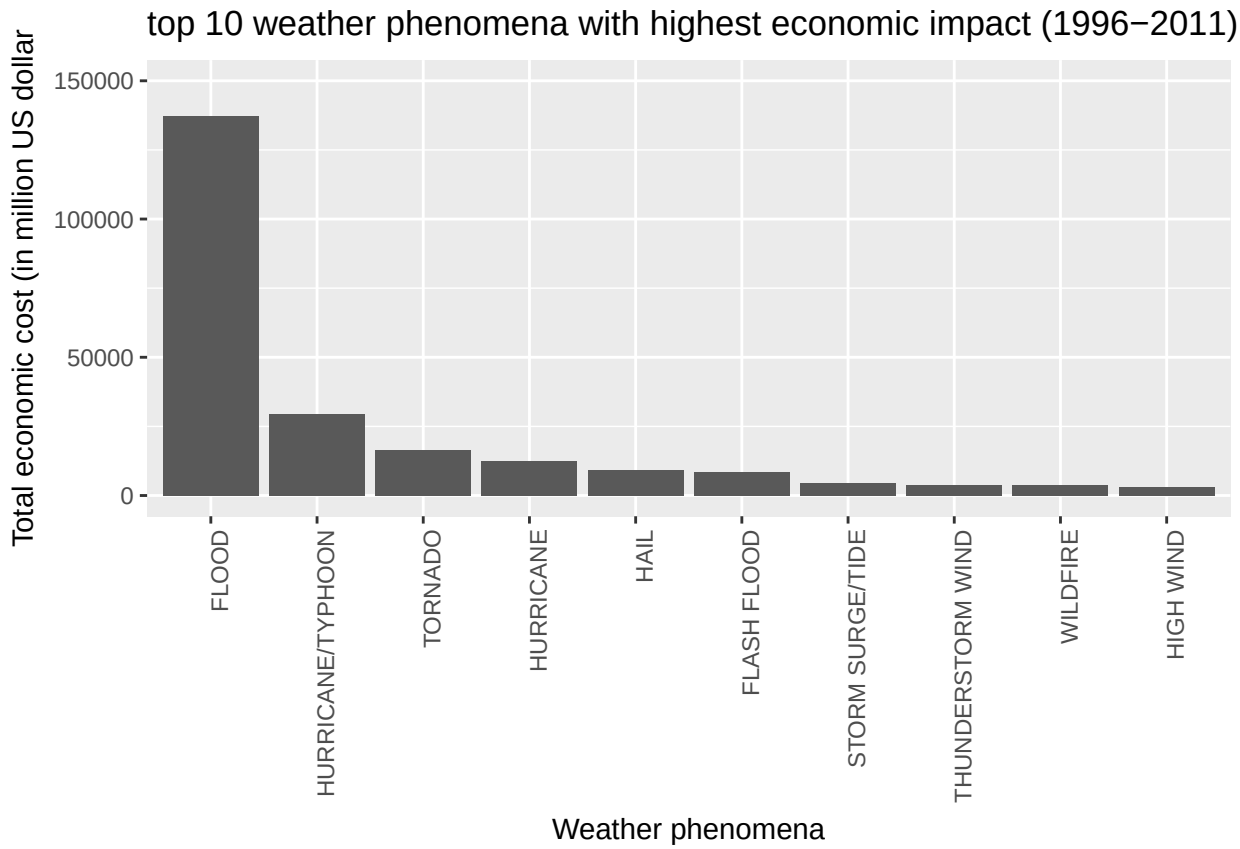
First we have to multiplie the base with the multiplier for both crops and properties. To do so we have to change the class of the exponential variables from charachter to numeric. We also express the total cost in millions by multiplying the cost with a factor 10^-6

```
stormnew$PROPDMGEXP<-as.numeric(stormnew$PROPDMGEXP)
stormnew$PROP_COST<-stormnew$PROPDMG*stormnew$PROPDMGEX
stormnew$CROPDMGEXP<-as.numeric(stormnew$CROPDMGEXP)
stormnew$CROP_COST<-stormnew$CROPDMG*stormnew$CROPDMGEX
stormnew$TOTAL_COST<-(stormnew$CROP_COST+stormnew$PROP_COST)*10^-6
```

Now we calculate the total economic cost for each event, order the data, select the top 10 and make a barplot. We also express the total cost in billions by multiplying the cost with a factor 10^-9

```
cost_event <- aggregate(TOTAL_COST ~ EVTYPE, data = stormnew, sum)
rank<-cost_event[order(cost_event$TOTAL_COST, decreasing=TRUE),][1:10,]

library(ggplot2)
z<-ggplot()+geom_bar(data=rank, aes(x=reorder(rank$EVTYPE,-rank$TOTAL_COST),
                                    y=rank$TOTAL_COST), stat="identity")
z<-z+theme(axis.text.x=element_text(angle=90, hjust=1))+ylim(0,150000)
z<-z+xlab("Weather phenomena")+ylab("Total economic cost (in million US dollar")+
        ggtitle("top 10 weather phenomena with highest economic impact (1996-2011)")+
        theme(axis.text.x=element_text(angle=90, hjust=1))
print(z)
```


top 10 weather phenomena with highest economic impact (1996−2011)
Total economic cost (in million US dollar
150000
100000
50000
0
FLOOD
HURRICANE/TYPHOON
TORNADO
HURRICANE
HAIL
FLASH FLOOD
STORM SURGE/TIDE
THUNDERSTORM WIND
WILDFIRE
HIGH WIND
Weather phenomena