Reproducible Research: Course Project 2

Severe weather events in the US and their impacts on health and economy

Synopsis

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern.

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database.

The data Storm Data record the severe weather event and the consequences from 1955 to 2011 across the US. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

There is also some documentation of the database available. Here are how some of the variables are constructed/defined.

- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQ

In this study, the obective is to find the type of events that causes the most important impact to public health and also has the greatest economic consequences. ## Data Processing Include the necessary packages

```
library(ggplot2)
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(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

Load data

```
if(!file.exists("repdata_data_StormData.csv.bz2")){
  Url<-"https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
  download.file(url =Url, "repdata_data_StormData.csv.bz2", method = "libcurl")
}
StormData<-read.csv("repdata_data_StormData.csv.bz2")
```

Show the dataframe's structure

```
str(StormData)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__    : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE   : Factor w/ 16335 levels "10/10/1954 0:00:00",..: 6523 6523 4213 11116 1426 1426 1462 28
##  $ BGN_TIME   : Factor w/ 3608 levels "000","0000","00:00:00 AM",..: 212 257 2645 1563 2524 3126 122 1
##  $ TIME_ZONE  : Factor w/ 22 levels "ADT","AKS","AST",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ COUNTY     : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: Factor w/ 29601 levels "","5NM E OF MACKINAC BRIDGE TO PRESQUE ISLE LT MI",..: 13513 1
##  $ STATE      : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ EVTYPE     : Factor w/ 985 levels "?","ABNORMALLY DRY",..: 830 830 830 830 830 830 830 830 830 830
##  $ BGN_RANGE  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI    : Factor w/ 35 levels "","E","Eas","EE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_LOCATI: Factor w/ 54429 levels "","?","(01R)AFB GNRY RNG AL",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_DATE   : Factor w/ 6663 levels "","10/10/1993 0:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_TIME   : Factor w/ 3647 levels "","?","0000",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI    : Factor w/ 24 levels "","E","ENE","ESE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_LOCATI: Factor w/ 34506 levels "","(0E4)PAYSON ARPT",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ LENGTH     : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH      : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F          : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES   : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG    : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17 17 17 17 17 ...
##  $ CROPDMG    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ WFO        : Factor w/ 542 levels "","2","43","9V9",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ STATEOFFIC: Factor w/ 250 levels "","ALABAMA, Central",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ZONENAMES  : Factor w/ 25112 levels "","
##  $ LATITUDE   : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE  : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS    : Factor w/ 436781 levels "","  ","  ","    ",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ REFNUM     : num  1 2 3 4 5 6 7 8 9 10 ...
```

- Relevant for the analysis are the date (**BGN_DATE**), event type (**EVTYPE**), counter for the health impact (**FATALITIES** and **INJURIES**), monetary impact on crop and property (**PROPDMG** and **CROPDMG**) as well as their corresponding exponents (**PROPDMGEXP** and **CROPDMGEXP**).
- According to the NOAA https://www.ncdc.noaa.gov/stormevents/details.jsp the full set of wheather events (48 event types) is available since 1996. Between 1950 and 1995 only a subset (Tornado,

Thunderstorm Wind and Hail) of these events is available in the storm database. In order to have a comparable basis for the analysis the dataset is limited to the observations between 1996 and 2011.

- The dataset contains a lot of observations without any information about health and/or economic damages. These observations are excluded from the analysis.
  Choose the useful variables in StormData

```
StormData<-select(StormData, BGN_DATE, EVTYPE, PROPDMG, PROPDMGEXP,
                  CROPDMG, CROPDMGEXP, FATALITIES, INJURIES)
StormData$BGN_DATE <- as.Date(StormData$BGN_DATE, "%m/%d/%Y")
StormData$Year<-year(StormData$BGN_DATE)
StormData<-filter(StormData, Year>=1996)
StormData <- filter(StormData, PROPDMG > 0 | CROPDMG > 0 | FATALITIES > 0 | INJURIES > 0)
```

The economical damages provided in the storm dataset require some adjustments. Each variable - **CROPDMG** and **PROPDMG** - comes with a separate exponent - **CROPDMGEXP** and **PROPDMGEXP**. First, the content of the exponent variables need to be converted into a proper coefficient.

```
table(StormData$PROPDMGEXP)
```

```
##
##             -      ?      +      0      1      2      3      4      5
##   8448      0      0      0      0      0      0      0      0      0
##      6      7      8      B      h      H      K      m      M
##      0      0      0     32      0      0 185474      0   7364
```

```
table(StormData$CROPDMGEXP)
```

```
##
##             ?      0      2      B      k      K      m      M
## 102767      0      0      0      2      0  96787      0   1762
```

There are some aboundant exponents, such as “h” and “H”, however the tables information shows all the lower case exponents have no data associated with. Here only the upper case exponents in **PRODMGEXP** and **CROPDMGEXP** are converted into corresponding coefficients:

“”, “?”, “+”, “-”: 1 “0”: 1
“1”: 10^1
“2”: 10^2
“3”: 10^3
“4”: 10^4
“5”: 10^5
“6”: 10^6
“7”: 10^7
“8”: 10^8
“9”: 10^9
“H”: 10^2
“K”: 10^3
“M”: 10^6
“B”: 10^9

```
StormData$CROPDMGCOEF[(StormData$CROPDMGEXP %in% c("","?","0"))] <-  1
StormData$CROPDMGCOEF[(StormData$CROPDMGEXP == "2")] <-  1e2
StormData$CROPDMGCOEF[(StormData$CROPDMGEXP == "K")] <-  1e3
StormData$CROPDMGCOEF[(StormData$CROPDMGEXP == "M")] <-  1e6
StormData$CROPDMGCOEF[(StormData$CROPDMGEXP == "B")] <-  1e9
```

```
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP %in% c("","-","?","+","0 "))]<-1
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="1")]<-1e1
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="2")]<-1e2
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="3")]<-1e3
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="4")]<-1e4
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="5")]<-1e5
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="6")]<-1e6
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="7")]<-1e7
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="8")]<-1e8
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="H")]<-1e2
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="K")]<-1e3
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="M")]<-1e6
StormData$PROPDMGCOEF[(StormData$PROPDMGEXP=="B")]<-1e9
```

Both variables the **INJURIES** and **FATALITIES** are added to form a new variable **Health_impact**. A similar aproach is used for the economic impact. Both crop and property damages are multiplied by their corresponding coefficients and added to form a new variable **Economic_cost**.

```
StormData <- mutate(StormData, Health_impact = FATALITIES + INJURIES)
StormData <- mutate(StormData, Economic_cost = PROPDMG * PROPDMGCOEF + CROPDMG * CROPDMGCOEF)
```

```
StormData$EVTYPE <- toupper(StormData$EVTYPE)
dim(table(StormData$EVTYPE))
```

```
## [1] 186
```

After converting the variable **EVTYPE** to uppercase, there are still 186 different event types listed. According to the NOAA there should be only 48. So there are a lot of duplicates.

Since this analysis is looking at the most harmful events, only part of the event types are cleaned. Therefore the health impact (**Health_impact**) is summed up per event type. Only the 10 event types with most **Health_impact** are kept.

```
HealthImpact_event<-summarise(group_by(StormData,EVTYPE),Health_impact=sum(Health_impact))
HealthImpact_event_principal<-arrange(HealthImpact_event, desc(Health_impact))[1:10,]
HealthImpact_event_principal
```

```
## # A tibble: 10 x 2
##    EVTYPE            Health_impact
##    <chr>                     <dbl>
##  1 TORNADO                   22178
##  2 EXCESSIVE HEAT             8188
##  3 FLOOD                      7172
##  4 LIGHTNING                  4792
##  5 TSTM WIND                  3870
##  6 FLASH FLOOD                2561
##  7 THUNDERSTORM WIND          1530
##  8 WINTER STORM               1483
##  9 HEAT                       1459
## 10 HURRICANE/TYPHOON          1339
```

There are two event types which are not compliant to the official types defined in the documentation. These are *TSTM WIND* and *HURRICANE/TYPHOON*.

```
HealthImpact_event_principal$EVTYPE[HealthImpact_event_principal$EVTYPE
                                    =="HURRICANE/TYPHOON"]<-"HURRICANE (TYPHOON)"
HealthImpact_event_principal$EVTYPE[HealthImpact_event_principal$EVTYPE
                                    =="TSTM WIND"]<-"THUNDERSTORM WIND"
```

The same procedure is been used to clean the event types for the most important economic impacts. After summing up the economic cost **Economic_cost**, 10 events with the greatest consequences are kept.

```
EconomicCost_event<-summarise(group_by(StormData,EVTYPE),Economic_cost=sum(Economic_cost))
EconomicCost_event_principal<-arrange(EconomicCost_event, desc(Economic_cost))[1:10,]
```

There are two event types which are not compliant to the official types defined in the documentation. These are *HURRICANE* and *STORM SURGE*.

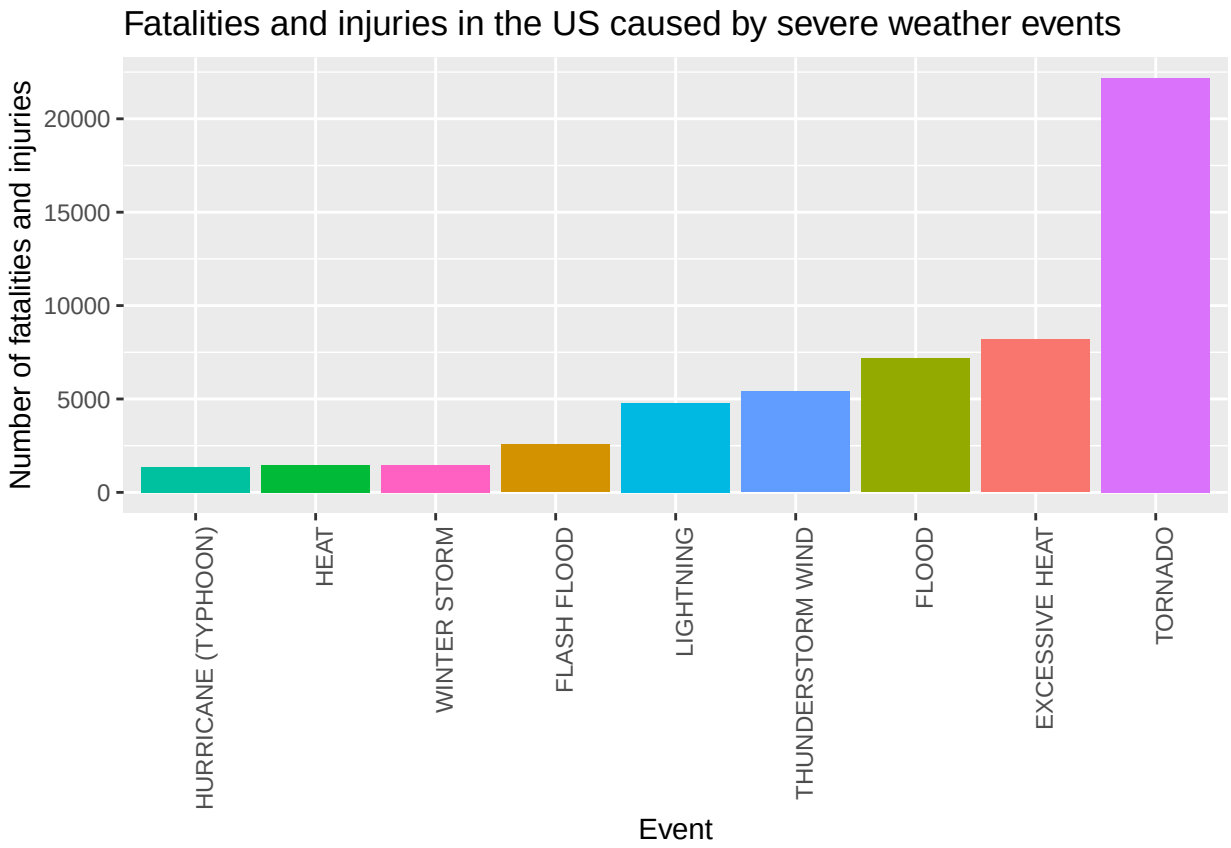
```
EconomicCost_event_principal$EVTYPE[EconomicCost_event_principal$EVTYPE
                                    =="HURRICANE"]<-"HURRICANE (TYPHOON)"
EconomicCost_event_principal$EVTYPE[EconomicCost_event_principal$EVTYPE
                                    =="STORM SURGE"]<-"STORM SURGE/TIDE"
```

Results

This is the plot for the types of events which are most harmful with respect to population health. The most harmful type of events is tornado.

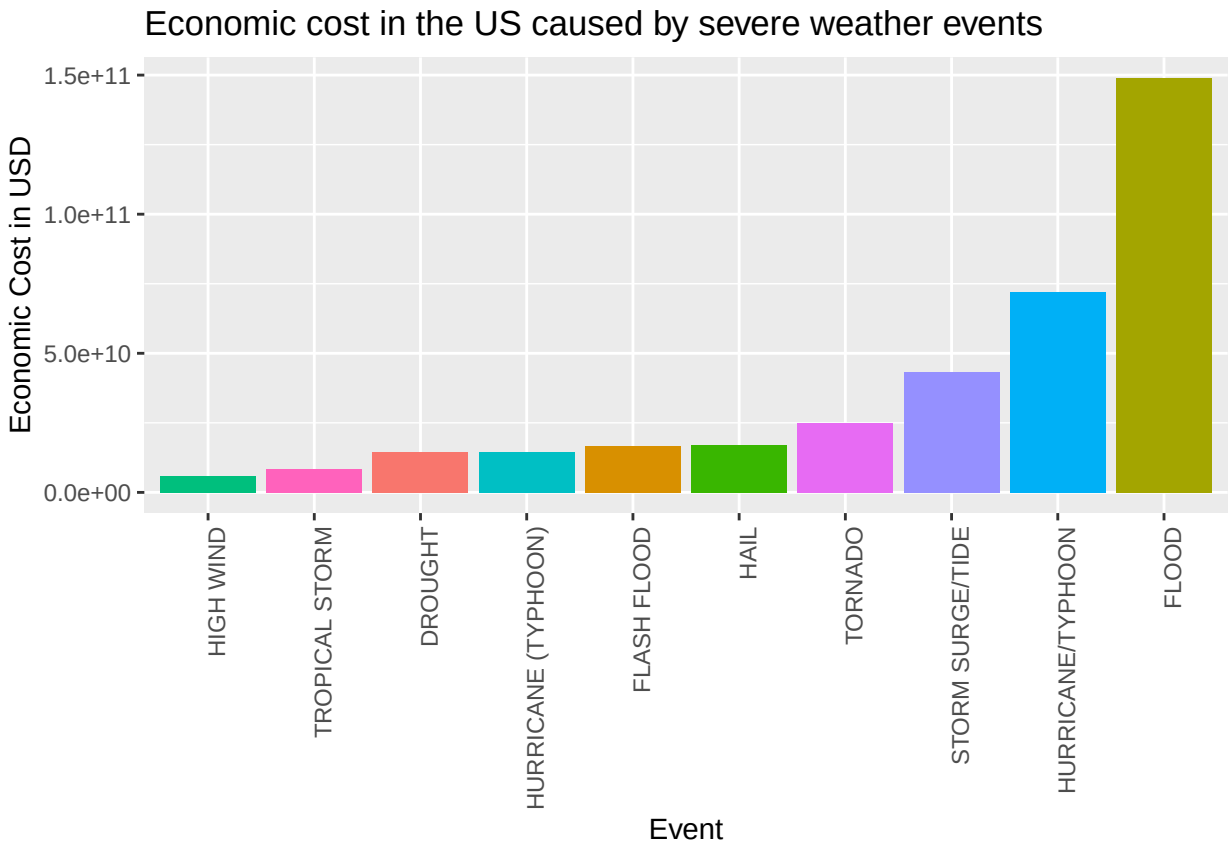
```
HealthImpact_event_final<-summarise(group_by(HealthImpact_event_principal,EVTYPE),
                                    Health_impact=sum(Health_impact))
HealthImpact_event_final<-arrange(HealthImpact_event_final,desc(Health_impact))

g1 <- ggplot(HealthImpact_event_final, aes(x=reorder(EVTYPE, Health_impact),
                                           y= Health_impact,fill=EVTYPE)) +
  geom_bar(stat="identity") +
  xlab("Event") + ylab("Number of fatalities and injuries") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),legend.position = "none") +
  ggtitle("Fatalities and injuries in the US caused by severe weather events")
g1
```



The following is the plot for the types of events which have the greatest economic consequences. Flood is the type that has the greatest economic consequences.

```
EconomicCost_event_final<-summarise(group_by(EconomicCost_event_principal,EVTYPE),
                                    Economic_cost=sum(Economic_cost))
EconomicCost_event_final<-arrange(EconomicCost_event_final,desc(Economic_cost))
g2 <- ggplot(EconomicCost_event_final, aes(x=reorder(EVTYPE,Economic_cost),
                                           y= Economic_cost,fill=EVTYPE)) +
  geom_bar(stat="identity") +
  xlab("Event") + ylab("Economic Cost in USD") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1),legend.position = "none") +
  ggtitle("Economic cost in the US caused by severe weather events")
g2
```


Economic cost in the US caused by severe weather events
Economic Cost in USD
1.5e+11
1.0e+11
5.0e+10
0.0e+00
HIGH WIND
TROPICAL STORM
DROUGHT
HURRICANE (TYPHOON)
FLASH FLOOD
HAIL
TORNADO
STORM SURGE/TIDE
HURRICANE/TYPHOON
FLOOD
Event