Reproducible Research - Course Project 2

Exploring the U.S. National Oceanic and Atmospheric Administration’s (NOAA) storm database - Health and Economic Impacts

Synopsis

This is a second course project for Reproducible Research course which is part of the Coursera’s Data Science Specialization.

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. This 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.

The analysis of the data shows that tornadoes, by far, have the greatest health impact as measured by the number of injuries and fatalities The analysis also shows that floods cause the greatest economic impact as measured by property damage and crop damage.

Data Processing

Load Libraries and prepare the R environment

I used these librarys in my analysis:

```
library(ggplot2) library(plyr) library(dplyr)
## ## Attaching package: 'dplyr'
## The following objects are masked from 'package:plyr': ## ## arrange,
count, desc, failwith, id, mutate, rename, summarise, ## summarize
```

```
## The following objects are masked from 'package:stats': ## ## filter, lag
## The following objects are masked from 'package:base': ## ## intersect,
setdiff, setequal, union
```

Data

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. You can download the file from the course web site:
storm data[47Mb]
There is also some documentation of the database available. Here you will find how some of the variables are constructed/defined.

- National Weather Service Storm Data Documentation

- National Climatic Data Center Storm Events FAQ

The events in the database start in the year 1950 and end in November 2011. 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.

Assignment

The basic goal of this assignment is to explore the NOAA Storm Database and answer the following basic questions about severe weather events.

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

Loading the data

The data was downloaded from the link above and saved on local computer (in setwd command one can replace loacal file path with path of folder where the data was downloaded). Then it was loaded on the R using the read.csv command. If object strom.data is already loaded, use that cached object insted of loading it each time the Rmd file is knitted.

```
if(!exists("storm.data")) { storm.data <-
read.csv(bzfile("repdata_data_StormData.csv.bz2"),header = TRUE) }
```

Examine the data set

In storm.data there is 37 columns (variables) and 902,297 rows (records).

```
dim(storm.data)
## [1] 902297 37
```

Examine the structure of the data

```
str(storm.data)
## '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 "1/1/1966 0:00:00",..:
6523 6523 4242 11116 2224 2224 2260 383 3980 3980 ... ## $ BGN_TIME : Factor
w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683 2584 3186 242 1683 3186
3186 ... ## $ 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 1873 4598 10592 4372 10094 1973 23873 24418 4598 ... ##
$ STATE : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ... ## $
EVTYPE : Factor w/ 985 levels " HIGH SURF ADVISORY",..: 834 834 834 834 834
834 834 834 834 834 ... ## $ BGN_RANGE : num 0 0 0 0 0 0 0 0 0 0 ... ## $
BGN_AZI : Factor w/ 35 levels ""," N"," NW",..: 1 1 1 1 1 1 1 1 1 1 ... ## $
BGN_LOCATI: Factor w/ 54429 levels "","- 1 N Albion",..: 1 1 1 1 1 1 1 1 1 1
... ## $ END_DATE : Factor w/ 6663 levels "","1/1/1993 0:00:00",..: 1 1 1 1 1
1 1 1 1 1 ... ## $ END_TIME : Factor w/ 3647 levels ""," 0900CST",..: 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 "","- .5 NNW",..: 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 ""," CI","$AC",..: 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 ""," "| __truncated__,..: 1 1 1
1 1 1 1 1 1 1 ... ## $ 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 "","-2 at Deer Park\n",..: 1 1 1 1 1 1 1 1 1 1 ... ## $ REFNUM : num 1
2 3 4 5 6 7 8 9 10 ...
```

Extracting variables of interest for analysis of weather impact on health and economy

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..)

Extract variables of interest from original data set:

```
vars <- c( "EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP",
"CROPDMG", "CROPDMGEXP") mydata <- storm.data[, vars]
```

Check the last few rows in data set (in firs years of recording there are many missing (NA) values):

```
tail(mydata)
## EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG ## 902292 WINTER
WEATHER 0 0 0 K 0 ## 902293 HIGH WIND 0 0 0 K 0 ## 902294 HIGH WIND 0 0 0 K 0
## 902295 HIGH WIND 0 0 0 K 0 ## 902296 BLIZZARD 0 0 0 K 0 ## 902297 HEAVY
SNOW 0 0 0 K 0 ## CROPDMGEXP ## 902292 K ## 902293 K ## 902294 K ## 902295 K
## 902296 K ## 902297 K
```

Checking for missing values

In every analysis we must the check number of missing values in variables.

Check for missing values in health variables - there is no NA's in the data.

```
sum(is.na(mydata$FATALITIES))
## [1] 0
sum(is.na(mydata$INJURIES))
## [1] 0
```

Check for missing values in economic variables for "size" of damage - there is no NA's in the data.

```
sum(is.na(mydata$PROPDMG))
## [1] 0
sum(is.na(mydata$CROPDMG))
## [1] 0
```

Check for missing values in economic variables for units damage - there is no NA's in the data.

```
sum(is.na(mydata$PROPDMGEXP))
## [1] 0
sum(is.na(mydata$CROPDMGEXP))
## [1] 0
```

Transforming extracted variables

Listing the first 10 event types that most appear in the data:

```
sort(table(mydata$EVTYPE), decreasing = TRUE)[1:10]
## ## HAIL TSTM WIND THUNDERSTORM WIND ## 288661 219940 82563 ## TORNADO
FLASH FLOOD FLOOD ## 60652 54277 25326 ## THUNDERSTORM WINDS HIGH WIND
LIGHTNING ## 20843 20212 15754 ## HEAVY SNOW ## 15708
```

We will group events like TUNDERSTORM WIND, TUNDERSTORM WINDS, HIGH WIND, etc. by containing the keyword 'WIND' as one event WIND. And we will transform other types of events in a similar way. New variable EVENTS is the transform variable of EVTYPE that have 10 different types of events: HEAT, FLOOD, etc., and type OTHER for events in which name the keyword is not found.

```
# create a new variable EVENT to transform variable EVTYPE in groups
mydata$EVENT <- "OTHER" # group by keyword in EVTYPE
mydata$EVENT[grep("HAIL", mydata$EVTYPE, ignore.case = TRUE)] <- "HAIL"
mydata$EVENT[grep("HEAT", mydata$EVTYPE, ignore.case = TRUE)] <- "HEAT"
mydata$EVENT[grep("FLOOD", mydata$EVTYPE, ignore.case = TRUE)] <- "FLOOD"
mydata$EVENT[grep("WIND", mydata$EVTYPE, ignore.case = TRUE)] <- "WIND"
mydata$EVENT[grep("STORM", mydata$EVTYPE, ignore.case = TRUE)] <- "STORM"
mydata$EVENT[grep("SNOW", mydata$EVTYPE, ignore.case = TRUE)] <- "SNOW"
mydata$EVENT[grep("TORNADO", mydata$EVTYPE, ignore.case = TRUE)] <- "TORNADO"
mydata$EVENT[grep("WINTER", mydata$EVTYPE, ignore.case = TRUE)] <- "WINTER"
mydata$EVENT[grep("RAIN", mydata$EVTYPE, ignore.case = TRUE)] <- "RAIN" #
listing the transformed event types sort(table(mydata$EVENT), decreasing =
TRUE)
## ## HAIL WIND STORM FLOOD TORNADO OTHER WINTER SNOW RAIN ## 289270 255362
113156 82686 60700 48970 19604 17660 12241 ## HEAT ## 2648
```

Checking the values for variables that represent units od dollars:

```
sort(table(mydata$PROPDMGEXP), decreasing = TRUE)[1:10]
## ## K M 0 B 5 1 2 ? m ## 465934 424665 11330 216 40 28 25 13 8 7
```

```
sort(table(mydata$CROPDMGEXP), decreasing = TRUE)[1:10]
## ## K M k 0 B ? 2 m <NA> ## 618413 281832 1994 21 19 9 7 1 1
```

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

New variable(s) is product of value of damage and dollar unit.

```
mydata$PROPDMGEXP <- as.character(mydata$PROPDMGEXP)
mydata$PROPDMGEXP[is.na(mydata$PROPDMGEXP)] <- 0 # NA's considered as dollars
mydata$PROPDMGEXP[!grepl("K|M|B", mydata$PROPDMGEXP, ignore.case = TRUE)] <-
0 # everything exept K,M,B is dollar mydata$PROPDMGEXP[grep("K",
mydata$PROPDMGEXP, ignore.case = TRUE)] <- "3" mydata$PROPDMGEXP[grep("M",
mydata$PROPDMGEXP, ignore.case = TRUE)] <- "6" mydata$PROPDMGEXP[grep("B",
mydata$PROPDMGEXP, ignore.case = TRUE)] <- "9" mydata$PROPDMGEXP <-
as.numeric(as.character(mydata$PROPDMGEXP)) mydata$property.damage <-
mydata$PROPDMG * 10^mydata$PROPDMGEXP mydata$CROPDMGEXP <-
as.character(mydata$CROPDMGEXP) mydata$CROPDMGEXP[is.na(mydata$CROPDMGEXP)]
<- 0 # NA's considered as dollars mydata$CROPDMGEXP[!grepl("K|M|B",
mydata$CROPDMGEXP, ignore.case = TRUE)] <- 0 # everything exept K,M,B is
dollar mydata$CROPDMGEXP[grep("K", mydata$CROPDMGEXP, ignore.case = TRUE)] <-
"3" mydata$CROPDMGEXP[grep("M", mydata$CROPDMGEXP, ignore.case = TRUE)] <-
"6" mydata$CROPDMGEXP[grep("B", mydata$CROPDMGEXP, ignore.case = TRUE)] <-
"9" mydata$CROPDMGEXP <- as.numeric(as.character(mydata$CROPDMGEXP))
mydata$crop.damage <- mydata$CROPDMG * 10^mydata$CROPDMGEXP
```

Print of first 10 values for property damage (in dollars) that most appear in the data:

```
sort(table(mydata$property.damage), decreasing = TRUE)[1:10]
## ## 0 5000 10000 1000 2000 25000 50000 3000 20000 15000 ## 663123 31731
21787 17544 17186 17104 13596 10364 9179 8617
```

Print of first 10 values for crop damage (in dollars) that most appear in the data:

```
sort(table(mydata$crop.damage), decreasing = TRUE)[1:10]
## ## 0 5000 10000 50000 1e+05 1000 2000 25000 20000 5e+05 ## 880198 4097
2349 1984 1233 956 951 830 758 721
```

Analysis

Aggregating events for public health variables

Table of public health problems by event type

```
# aggregate FATALITIES and INJURIES by type of EVENT
agg.fatalites.and.injuries <- ddply(mydata, .(EVENT), summarize, Total =
sum(FATALITIES + INJURIES, na.rm = TRUE)) agg.fatalites.and.injuries$type <-
"fatalities and injuries" # aggregate FATALITIES by type of EVENT
agg.fatalities <- ddply(mydata, .(EVENT), summarize, Total = sum(FATALITIES,
na.rm = TRUE)) agg.fatalities$type <- "fatalities" # aggregate INJURIES by
type of EVENT agg.injuries <- ddply(mydata, .(EVENT), summarize, Total =
sum(INJURIES, na.rm = TRUE)) agg.injuries$type <- "injuries" # combine all
agg.health <- rbind(agg.fatalities, agg.injuries) health.by.event <- join
(agg.fatalities, agg.injuries, by="EVENT", type="inner") health.by.event
## EVENT Total type Total type ## 1 FLOOD 1524 fatalities 8602 injuries ## 2
HAIL 15 fatalities 1371 injuries ## 3 HEAT 3138 fatalities 9224 injuries ## 4
OTHER 2626 fatalities 12224 injuries ## 5 RAIN 114 fatalities 305 injuries ##
6 SNOW 164 fatalities 1164 injuries ## 7 STORM 416 fatalities 5339 injuries
## 8 TORNADO 5661 fatalities 91407 injuries ## 9 WIND 1209 fatalities 9001
injuries ## 10 WINTER 278 fatalities 1891 injuries
```

Aggregating events for economic variables

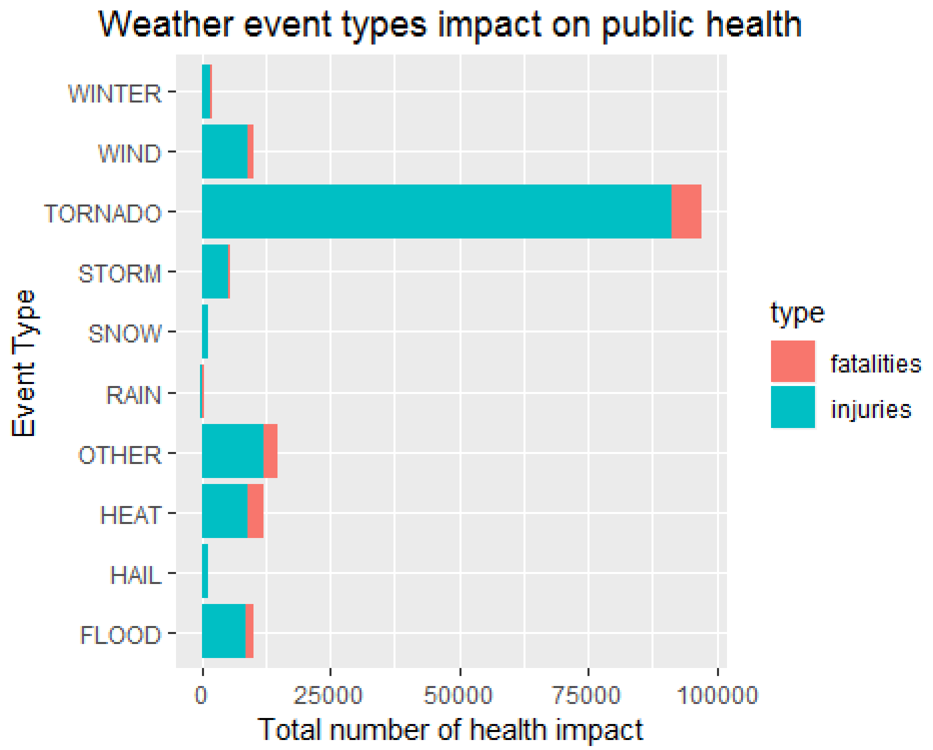
```
# aggregate PropDamage and CropDamage by type of EVENT
agg.propdmg.and.cropdmg <- ddply(mydata, .(EVENT), summarize, Total =
sum(property.damage + crop.damage, na.rm = TRUE))
agg.propdmg.and.cropdmg$type <- "property and crop damage" # aggregate
PropDamage by type of EVENT agg.prop <- ddply(mydata, .(EVENT), summarize,
Total = sum(property.damage, na.rm = TRUE))
```

```
agg.prop$type <- "property" # aggregate INJURIES by type of EVENT agg.crop <-
ddply(mydata, .(EVENT), summarize, Total = sum(crop.damage, na.rm = TRUE))
agg.crop$type <- "crop" # combine all agg.economic <- rbind(agg.prop,
agg.crop) economic.by.event <- join (agg.prop, agg.crop, by="EVENT",
type="inner") economic.by.event
## EVENT Total type Total type ## 1 FLOOD 167502193929 property 12266906100
crop ## 2 HAIL 15733043048 property 3046837473 crop ## 3 HEAT 20325750
property 904469280 crop ## 4 OTHER 97246712337 property 23588880870 crop ## 5
RAIN 3270230192 property 919315800 crop ## 6 SNOW 1024169752 property
134683100 crop ## 7 STORM 66304415393 property 6374474888 crop ## 8 TORNADO
58593098029 property 417461520 crop ## 9 WIND 10847166618 property 1403719150
crop ## 10 WINTER 6777295251 property 47444000 crop
```

Results

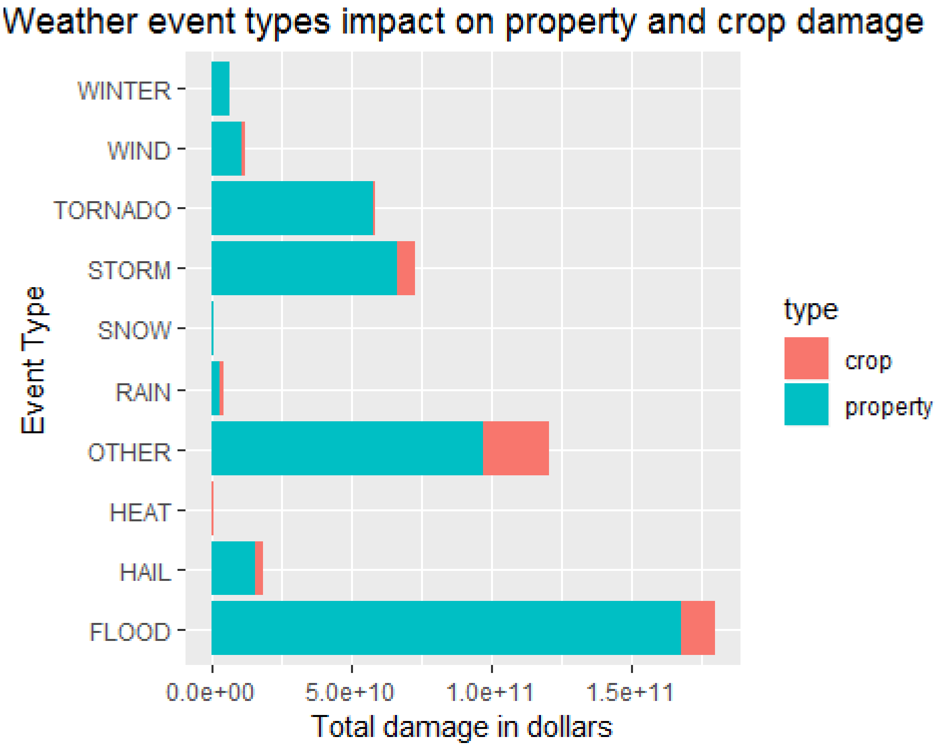
Across the United States, which types of events are most harmful with respect to population health?

```
# transform EVENT to factor variable for health variables agg.health$EVENT <-
as.factor(agg.health$EVENT) # plot FATALITIES and INJURIES by EVENT
health.plot <- ggplot(agg.health, aes(x = EVENT, y = Total, fill = type)) +
geom_bar(stat = "identity") + coord_flip() + xlab("Event Type") + ylab("Total
number of health impact") + ggtitle("Weather event types impact on public
health") + theme(plot.title = element_text(hjust = 0.5)) print(health.plot)
```



The most harmful weather event for health (in number of total fatalites and injuries) is, by far, a tornado. ### Across the United States, which types of events have the greatest economic consequences?

```
# # transform EVENT to factor variable for economic variables
agg.economic$EVENT <- as.factor(agg.economic$EVENT) # plot PROPERTY damage
and CROP damage by EVENT economic.plot <- ggplot(agg.economic, aes(x = EVENT,
y = Total, fill = type)) + geom_bar(stat = "identity") + coord_flip() +
xlab("Event Type") + ylab("Total damage in dollars") + ggtitle("Weather event
types impact on property and crop damage") + theme(plot.title =
element_text(hjust = 0.5)) print(economic.plot)
```



The most devastating weather event with the greatest economic cosequences (to property and crops) is a flood.