Impact of Weather Events on Health and Economic

Introduction

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.

About the 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. 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.

About the assignment

The basic goal of this assignment is to explore the NOAA Storm Database and answer some basic questions about severe weather events. You must use the database to answer the questions below and show the code for your entire analysis. Your analysis can consist of tables, figures, or other summaries. You may use any R package you want to support your analysis.

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?

Notes

This study was done using the following tools:

```
-a 64-bit Windows 7 machine with 4 cores.
-R language was R version 3.3.2 (2016-10-31), RStudio
```

Because of some account trouble, Rpubs is not used. The full project can be found on Github at https://github.com/michael-gm/repoducible.

Data Processing

Loading Libraries

```
library(stringr)
library(lubridate)
library(sqldf)
library(ggplot2)
library(reshape2)
library(gridExtra)
```

Loading data

```
dseturl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
dsetzip <- "data/StormData.csv.bz2"
dsetrds <- "data/StormData.RDS"

if (!file.exists(dsetzip)) {
    download.file(url = dseturl,
                  destfile = dsetzip,
                  method = "curl")
}

RDSloaded <- FALSE
if (!file.exists(dsetrds)) {
    d <- read.csv(file = bzfile(dsetzip), strip.white = TRUE)
} else {
    d <- readRDS(dsetrds)
    RDSloaded <- TRUE
}
```

Cleaning the data

We want to find out which events are the worst we have to use the property damage (`PROPDMG`) and crop damage (`CROPDMG`) values. These are simple integers which must be multiplied by an exponent given in another field (`PROPDMGEXP` and `CROPDMGEXP`. Unfortunately, not all of the exponent values are valid, some appear to be rounding.

We allowed only the following values for exponent: * H hundred (x100) * K thousand (x1,000) * M million (x1,000,000) * B billion (x1,000,000,000)

```
if(!RDSloaded) {
    calcUSD <- function(dmg, dmgexp) dmg * switch(toupper(dmgexp), H=100, K=1000,
                                                  M=1000000, B=1000000000, 1)

    d$pdmgUSD <- mapply(calcUSD, d$PROPDMG, d$PROPDMGEXP)
    d$cdmgUSD <- mapply(calcUSD, d$CROPDMG, d$CROPDMGEXP)
}
```

We convert dates to POSIXct format.

```
if(!RDSloaded)
    d$BEGIN_UTC <- mdy(str_extract(d$BGN_DATE, "[^ ]+"))
```

The data is in some cases in a badly quality. There are columsn which has many different events, but they mean almost the same, e.g. abbreviated or full text of the same event. (`thunderstorm`, `gusty thunderstorm`

`wind`, `gusty wind/rain`, `marine tstm wind`). We are going to categorize the most impactful events by looking for common words and abbreviations in a relative handful of weather categories.

```
if (!RDSloaded) {
    generateEvent <- function(evt) {
        evt <- tolower(evt)
        ifelse(grepl("lightning", evt), "lightning",
        ifelse(grepl("hail", evt), "hail",
        ifelse(grepl("rain|flood|wet|fld", evt), "rain",
        ifelse(grepl("snow|winter|wintry|blizzard|sleet|cold|ice|
                     freeze|avalanche|icy", evt), "winter",
        ifelse(grepl("thunder|tstm|tornado|wind|hurricane|funnel|
                     tropical +storm", evt), "wind",
        ifelse(grepl("fire", evt), "fire",
        ifelse(grepl("fog|visibility|dark|dust", evt),
               "low visibility",
        ifelse(grepl("surf|surge|tide|tsunami|current", evt),
               "ocean surge",
        ifelse(grepl("heat|high +temp|record +temp|warm|dry", evt),
               "heat",
        ifelse(grepl("volcan", evt), "volcanic activity",
               "uncategorized"
        ))))))))))
    }
    d$weatherCategory <- mapply(generateEvent, d$EVTYPE)
}
```

For purposes of this study, USA is defined as the 50 states in the continental US, plus District of Columbia, Hawaii and Alaska. territories, protectorates, and military regions are excluded.

```
if (!RDSloaded)
    d$isUSA <- mapply( function(st) st %in% state.abb, d$STATE )
```

In the interest of performance, we will save the data frame as an R data set.

```
if (!RDSloaded) {
    saveRDS(d, file=dsetrds)
    d <- readRDS(dsetrds)
    RDSloaded <- TRUE
}
## subset to only USA data
d <- d[d$isUSA == TRUE,]
```

Results

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

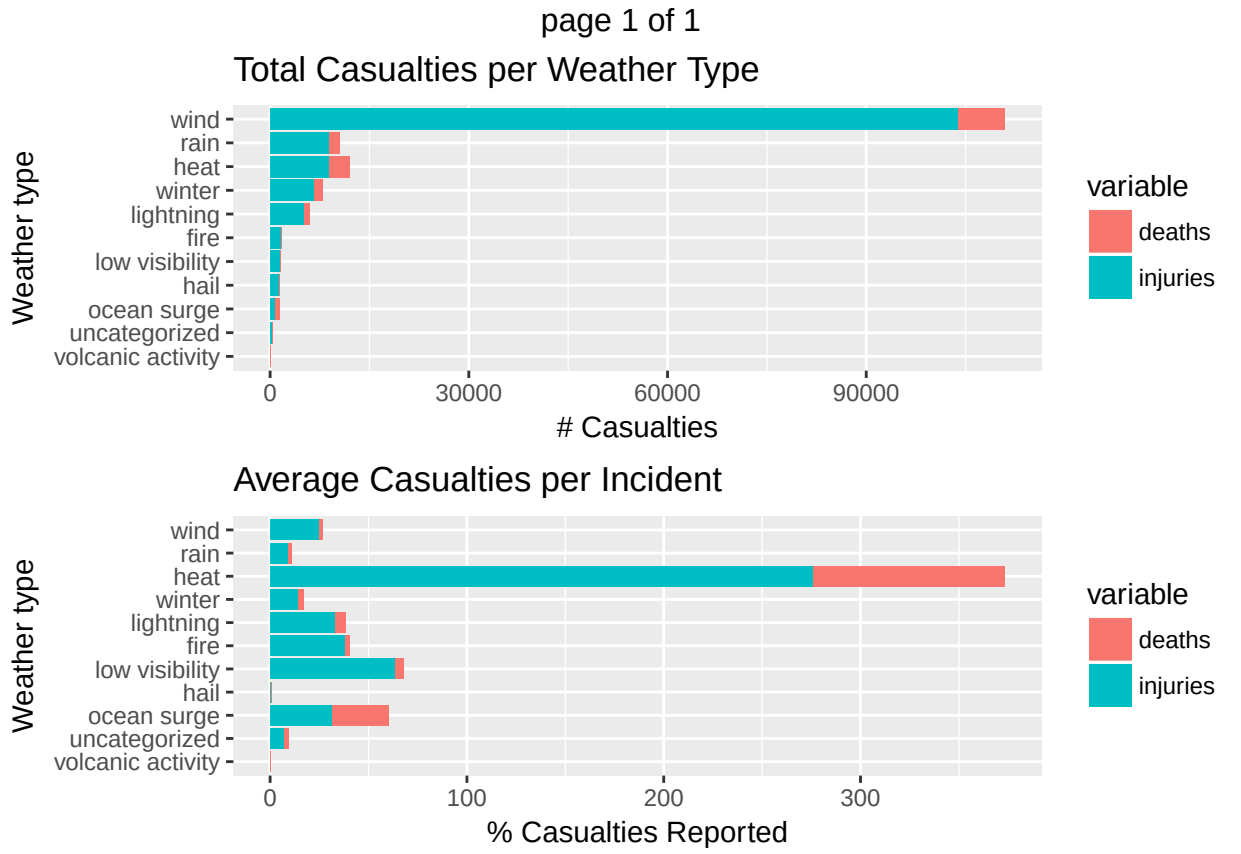
Refer to the result of the query below, which groups US mortalities and injuries by weather category.

```
harm <- sqldf("select sum(FATALITIES) as deaths, sum(INJURIES)
               as injuries, weatherCategory,count(*) as sumrecs
              from d group by weatherCategory ")

harm$weatherCategory <- factor(harm$weatherCategory,
```

```
        levels=harm[order(harm$injuries), "weatherCategory"])

hdat <- melt(harm, id.vars=c("weatherCategory", "sumrecs"),
             measure.vars=c("deaths", "injuries"))
hdat <- sqldf("select *,(value/sumrecs)*100 as pctPerEvent
              from hdat")
plot1 <- ggplot(hdat, aes(x=weatherCategory, y=value,
            fill=variable)) + geom_bar(stat="identity") +
    coord_flip() + ggtitle("Total Casualties per Weather Type") +
    xlab("Weather type") + ylab("# Casualties")
plot2 <- ggplot(hdat, aes(x=weatherCategory, y=pctPerEvent,
    fill=variable)) + geom_bar(stat="identity") + coord_flip() +
    ggtitle("Average Casualties per Incident") +
    xlab("Weather type") + ylab("% Casualties Reported")
marrangeGrob(list(plot1, plot2), nrow = 2, ncol = 1)
```

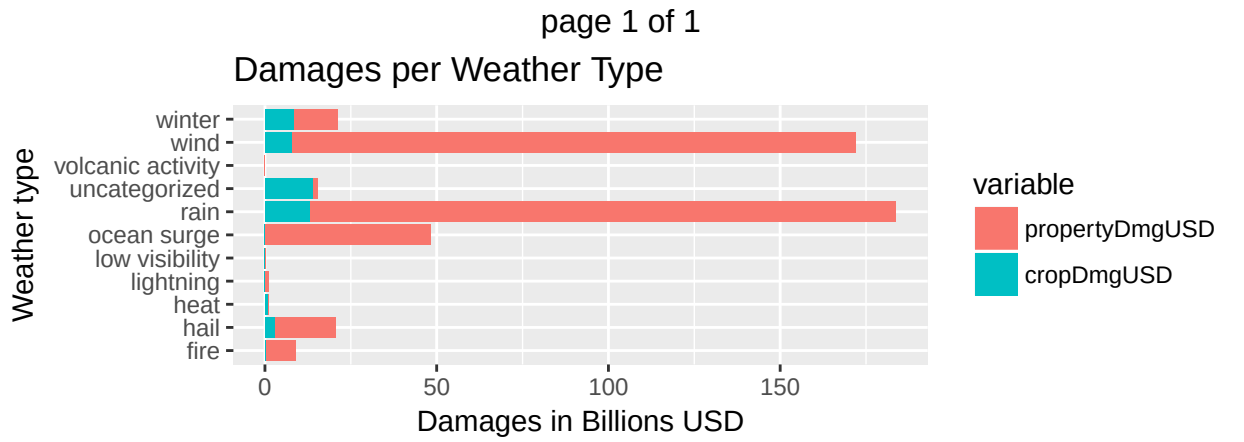


Wind events – including tornadoes and hurricanes – have the highest impact on health in terms of absolute numbers reported. Heat events – including fires and heat waves – show the highest percentage of casualties per event.

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

```
crop <- sqldf("select sum(pdmgUSD) as propertyDmgUSD,
            sum(cdmgUSD) as cropDmgUSD, count(*) as
```

```
            sumrecs, weatherCategory from d group by
            weatherCategory")
crop <- sqldf("select *, propertyDmgUSD +
              cropDmgUSD as totalCost from crop")
## create long form on crop vs property damage columns
cdat <- melt(crop, id.vars=c("weatherCategory",
            "sumrecs", "totalCost"), measure.vars=
               c("propertyDmgUSD", "cropDmgUSD"))
cdat <- sqldf("select *, (value/sumrecs) as
              costPerEvent from cdat")
plot3 <- ggplot(cdat, aes(x=weatherCategory,
            y=value/1000000000, fill=variable)) +
    geom_bar(stat="identity") + coord_flip() +
    ggtitle("Damages per Weather Type") +
    xlab("Weather type") +
    ylab("Damages in Billions USD")
marrangeGrob(list(plot3), nrow=2, ncol=1)
```



Rain and wind events are the most costly weather types, both in terms of property and damage to crops. There is a very significant crop damage cost in the "uncategorized" weather events. This bears examination, and possibly some re-evaluation of the categorization used in this study.