Title: Reproducible Research - Course Project 2

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.

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 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

Consider writing your report as if it were to be read by a government or municipal manager who might be responsible for preparing for severe weather events and will need to prioritize resources for different types of events. However, there is no need to make any specific recommendations in your report.

Your data analysis must address the following 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?

Synopsis

In this report we did some analysis to answer the above questions.

According to our analysis the top 10 events that are most harmful with respect to population health are:

- TORNADO, EXCESSIVE HEAT, FLASH FLOOD, HEAT, LIGHTNING, TSTM WIND, FLOOD, RIP CURRENT, HIGH WIND, AVALANCHE caused most fatalities from 1950 to 2011.
- TORNADO, TSTM WIND, FLOOD, EXCESSIVE HEAT, LIGHTNING, HEAT, ICE STORM, FLASH FLOOD, THUNDERSTORM WIND, HAIL caused most injuries from 1950 to 2011.

The following 10 to events have the greatest economic consequences:

- FLOOD, HURRICANE/TYPHOON, TORNADO, STORM SURGE, FLASH FLOOD, HAIL, HURRICANE, TROPICAL STORM, WINTER STORM, HIGH WIND have most property damage from 1950 to 2011.
- DROUGHT, FLOOD, RIVER FLOOD, ICE STORM, HAIL, HURRICANE, HURRICANE/TYPHOON, FLASH FLOOD, EXTREME COLD, FROST/FREEZE have most crop damage from 1950 to 2011.

Data Processing

```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.2.3
```

Let's first download and unzip the data set if we have not already done it before.

Then we read the data set.

```
url = "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
file = "repdata_data_StormData.csv.bz2"
if(!file.exists(file)) download.file(url = url, destfile = file)
df  <- read.csv("repdata_data_StormData.csv.bz2")
```

Now let's look at the summary of the data set.

```
str(df)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ BGN_TIME  : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE : chr  "CST" "CST" "CST" "CST" ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ 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   : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ 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: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ 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   : chr  "" "" "" "" ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

There are approximately 1 million observation and 37 variables in this data set. Not all of the variables are useful for our analysis. Therefore we only pick the ones that we need for our analysis to answer the questions.

```
selected_df <- df[, c("EVTYPE", "PROPDMG", "PROPDMGEXP", "CROPDMG",
                      "CROPDMGEXP", "FATALITIES", "INJURIES")]
```

Then we do some data cleaning to prepare the data for finla analysis.

```
#table(selected_df$PROPDMGEXP)
selected_df[selected_df$PROPDMGEXP %in%
                    c("0", "1", "2", "3", "4", "5", "6", "7", "8"), ]$PROPDMGEXP <- 10
selected_df[selected_df$PROPDMGEXP %in% c("", "-", "?"), ]$PROPDMGEXP <- 0
selected_df[selected_df$PROPDMGEXP %in% c("+"), ]$PROPDMGEXP <- 1
selected_df[selected_df$PROPDMGEXP %in% c("h", "H"), ]$PROPDMGEXP <- 100
selected_df[selected_df$PROPDMGEXP %in% c("K", "k"), ]$PROPDMGEXP <- 1000
selected_df[selected_df$PROPDMGEXP %in% c("M", "m"), ]$PROPDMGEXP <- 1000000
selected_df[selected_df$PROPDMGEXP %in% c("B", "b"), ]$PROPDMGEXP <- 1000000000
selected_df$PROPDMGEXP <- as.numeric(selected_df$PROPDMGEXP)
#table(selected_df$PROPDMGEXP)

#table(selected_df$CROPDMGEXP)
selected_df[selected_df$CROPDMGEXP == "0" | selected_df$CROPDMGEXP == "2", ]$CROPDMGEXP <- 10
selected_df[selected_df$CROPDMGEXP == "" | selected_df$CROPDMGEXP == "?", ]$CROPDMGEXP <- 0
selected_df[selected_df$CROPDMGEXP == "K" | selected_df$CROPDMGEXP == "k", ]$CROPDMGEXP <- 1000
selected_df[selected_df$CROPDMGEXP == "M" | selected_df$CROPDMGEXP == "m", ]$CROPDMGEXP <- 1000000
selected_df[selected_df$CROPDMGEXP == "B", ]$CROPDMGEXP <- 1000000000
selected_df$CROPDMGEXP <- as.numeric(selected_df$CROPDMGEXP)
#table(selected_df$CROPDMGEXP)
```

```
selected_df$FINALCROPDMG <- selected_df$CROPDMG*selected_df$CROPDMGEXP
selected_df$FINALPROPDMG <- selected_df$PROPDMG*selected_df$PROPDMGEXP
```

Results

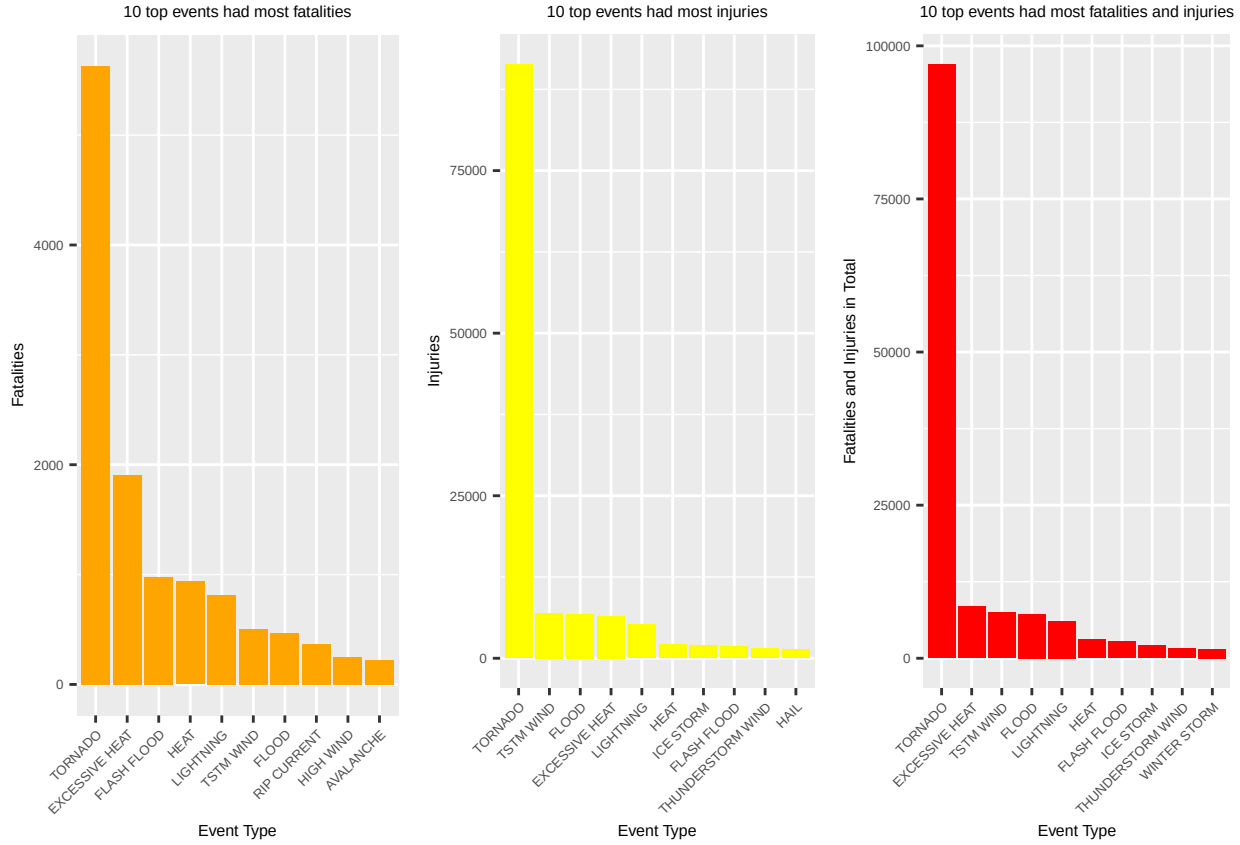
First we show the 10 top events that are most harmful with respect to population health.

```
total_injuries <- aggregate(INJURIES ~ EVTYPE, data=selected_df, FUN = sum)

total_injuries <- total_injuries[order(total_injuries$INJURIES,
                                       decreasing = TRUE), ][1:10,]

total_fatalities <- aggregate(FATALITIES ~ EVTYPE, data=selected_df , FUN = sum)
total_fatalities <- total_fatalities[order(total_fatalities$FATALITIES,
                                   decreasing = TRUE), ][1:10, ]

both_injfat <- aggregate(FATALITIES + INJURIES ~ EVTYPE, data=selected_df, FUN = sum)
both_injfat <- both_injfat[order(both_injfat$`FATALITIES + INJURIES`,
                                    decreasing=TRUE), ][1:10, ]

both_injfat$BOTH <- both_injfat$`FATALITIES + INJURIES`
both_injfat$`FATALITIES + INJURIES` <- NULL
```

The following charts display the 10 top events that caused most fatalities, injuries and both fatalities and injuries together accordingly.

```
par(mfrow = c(1, 3))

total_fatalities$EVTYPE <- reorder(total_fatalities$EVTYPE, -total_fatalities$FATALITIES)
plot1 <- ggplot(data=total_fatalities, aes(x=EVTYPE, y=FATALITIES)) +
        geom_bar(stat = "identity", fill = "orange") +
        labs(title = "10 top events had most fatalities",
             x = "Event Type",
             y = "Fatalities") +
        theme(text = element_text(size = 6),
              axis.text.x = element_text(angle = 45, hjust = 1),
              plot.title = element_text(size = 6, hjust = 0.5))

total_injuries$EVTYPE <- reorder(total_injuries$EVTYPE, -total_injuries$INJURIES)
plot2 <- ggplot(data=total_injuries, aes(x=EVTYPE, y=INJURIES)) +
        geom_bar(stat = "identity", fill = "yellow") +
        labs(title = "10 top events had most injuries",
             x = "Event Type",
             y = "Injuries") +
        theme(text = element_text(size = 6),
              axis.text.x = element_text(angle = 45, hjust = 1),
              plot.title = element_text(size = 6, hjust = 0.5))

both_injfat$EVTYPE <- reorder(both_injfat$EVTYPE, -both_injfat$BOTH)
plot3 <- ggplot(data=both_injfat, aes(x=EVTYPE, y=BOTH)) +
        geom_bar(stat = "identity", fill = "red") +
        labs(title = "10 top events had most fatalities and injuries",
```

```
             x = "Event Type",
             y = "Fatalities and Injuries in Total") +
        theme(text = element_text(size = 6),
              axis.text.x = element_text(angle = 45, hjust = 1),
              plot.title = element_text(size = 6, hjust = 0.5))

grid.arrange(plot1, plot2, plot3, nrow = 1)
```



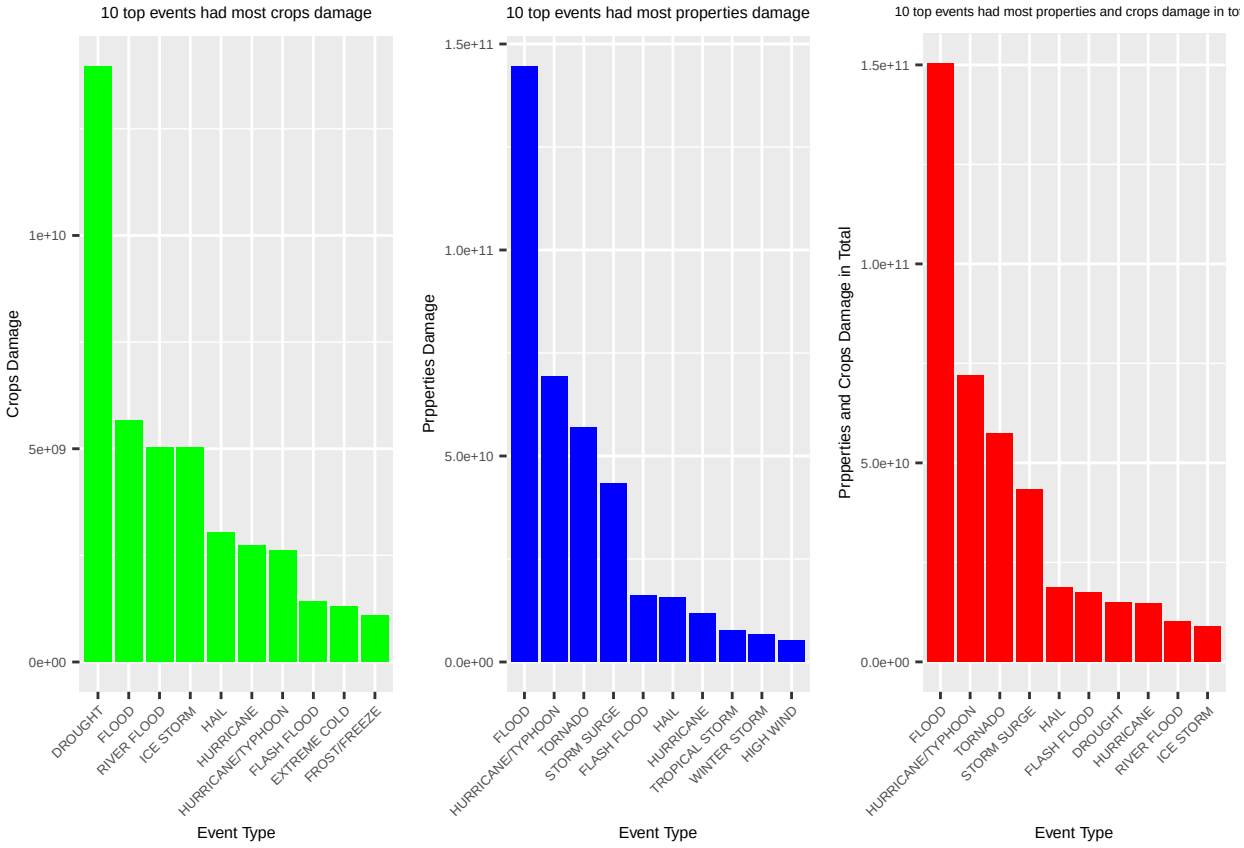
Then we show the 10 top events that have the greatest economic consequences.

```
total_cropdmg <- aggregate(FINALCROPDMG ~ EVTYPE, data=selected_df, FUN=sum)
total_cropdmg <- total_cropdmg[order(total_cropdmg$FINALCROPDMG, decreasing = TRUE), ]
total_cropdmg <- total_cropdmg[1:10, ]

total_propdmg <- aggregate(FINALPROPDMG ~ EVTYPE, data=selected_df, FUN = sum)
total_propdmg <- total_propdmg[order(total_propdmg$FINALPROPDMG, decreasing = TRUE), ]
total_propdmg <- total_propdmg[1:10, ]

propcropdmg <- aggregate(FINALPROPDMG + FINALCROPDMG ~ EVTYPE,
                         data=selected_df, FUN=sum)
propcropdmg$BOTH <- propcropdmg$`FINALPROPDMG + FINALCROPDMG`
propcropdmg$`FINALPROPDMG + FINALCROPDMG` <- NULL

propcropdmg <- propcropdmg[order(propcropdmg$BOTH, decreasing = TRUE), ][1:10,]
```

The following charts display the 10 top events that caused most properties, crops and properties and crops

damage together accordingly.

```
par(mfrow = c(1, 3))
total_cropdmg$EVTYPE <- reorder(total_cropdmg$EVTYPE, -total_cropdmg$FINALCROPDMG)
plot1 <- ggplot(data=total_cropdmg, aes(x=EVTYPE, y=FINALCROPDMG)) +
        geom_bar(stat = "identity", fill="green") +
        labs(title = "10 top events had most crops damage",
             x = "Event Type",
             y = "Crops Damage") +
        theme(text = element_text(size = 6),
              axis.text.x = element_text(angle = 45, hjust = 1),
              plot.title = element_text(size = 6, hjust = 0.5))

total_propdmg$EVTYPE <- reorder(total_propdmg$EVTYPE, -total_propdmg$FINALPROPDMG)
plot2 <- ggplot(data=total_propdmg, aes(x=EVTYPE, y=FINALPROPDMG)) +
        geom_bar(stat = "identity", fill="blue") +
         labs(title = "10 top events had most properties damage",
             x = "Event Type",
             y = "Prpperties Damage") +
        theme(text = element_text(size = 6),
              axis.text.x = element_text(angle = 45, hjust = 1),
              plot.title = element_text(size = 6, hjust = 0.5))

propcropdmg$EVTYPE <- reorder(propcropdmg$EVTYPE, -propcropdmg$BOTH)
plot3 <- ggplot(data=propcropdmg, aes(x=EVTYPE, y=BOTH)) +
        geom_bar(stat = "identity", fill="red") +
        labs(title = "10 top events had most properties and crops damage in total",
             x = "Event Type",
             y = "Prpperties and Crops Damage in Total") +
        theme(text = element_text(size = 6),
              axis.text.x = element_text(angle = 45, hjust = 1),
              plot.title = element_text(size = 5, hjust = 0.5))

grid.arrange(plot1, plot2, plot3, nrow = 1)
```



Reference: https://www.coursera.org/learn/reproducible-research/peer/OMZ37/course-project-2