$Project2 : ReproducibleResearch$

Severe Weather And Its Impact On Public Health and Economy

Abstract

In this report, the aim is to analyze the impact of disparate and connected weather events on public health and the economy. Reliance is on the storm database of the U.S. National Oceanic and Atmospheric Administration's (NOAA) from 1950 to 2011. Estimates of fatalities, injuries, property and crop damage were used to provide assistance with which types of event are most harmful to the population health and what the economic impact might be. Excessive heat and tornados were found to be most harmful with respect to population health, while floods, drought, and hurricanes have the greatest economic consequences.

Introduction

Storms and other catastrophic weather events cause both public health and economic problems for communities of various dimensions. Many storm events cause personal injuries or, in extreme cases, fatalities, with property damage a secondary concern. Preventing such outcomes to the extent possible is a major issue.

This project involves exploring the NOAA storm database. The utility allows tracking characteristics of storms events, including time and location of estimates of fatalities, injuries, and property damage.

Storm Data is available as a compressed, downloadable file, of 47Mb (megabytes). A document pairs with the data to provide required knowledge about the use and partitioning of the data file. For assistance, a Frequently Asked Qustions (FAQ) web page is available.

Database entry, initiated during April 1950, ended in November 2011. Fewer events were recorded for years prior to 1995. This was most likely due to a lack of diligent record-keeping. Dates after this time period were considered as more complete and used in the report.

The report is seperated into several sections - Abstract, Introduction, Data Processing, Methods, Results and Conclusions. The Methods section encompases the data analysis and presents two primary questions: 1. regarding the United States, which types of events are most harmful with respect to population health, and 2. which events have the greatest economic consequences?

To assist in answering these issues, figures are included to illustrate significant points in the report. Recommendations are not suggested, but the presentation of statistically inferred data analysis is significant. For transparency, the data source is available from NOAA and R-code can be acquired in its entirety from the author.

Initial settings and loading of required libraries

```
library(R.utils)
library(ggplot2)
library(plyr)
require(gridExtra)

warning = FALSE
echo = TRUE            # Globally, make code visible
options(scipen = 2)    # No scientific notation, but 2 dp
```

Data Processing

Redirect R-code and data file store paths. Check for data accessibility, if the data has been previously extracted, we do not need to load it again. Otherwise, download and unzip the file.

```
setwd("K:/COURSES/JHU_DataScience/ReproducibleResearch/Project_2")

#is there a stormData file, it not then download at NOAA's website and unzip displaying a progress bar
if (!"NOAA_StormData.csv.bz2" %in% dir("data"))
{
  message("PLZ WAIT, downloading ... this may take several minutes")
  download.file("http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",
                destfile="data/NOAA_StormData.csv.bz2",
                quiet=FALSE)
}

#unzip the data file to 'NOAA_StormData.csv'
if (!"NOAA_StormData.csv" %in% dir("data"))
{
    message("unzipping data file")
    bunzip2("NOAA_StormData.csv.bz2", overwrite=FALSE, remove=FALSE)
}
```

check for a ‘NOAA_StormData’ datafile loaded, then read if available

```
if (!"stormData" %in% ls())
{
    message("reading NOAA data file")
    stormData <- read.csv("data/NOAA_StormData.csv", sep = ",", stringsAsFactors=FALSE)
}
```

```
## reading NOAA data file
```

```
message("READY to process")
```

```
## READY to process
```

Inspect a small section of the ‘RAW’ data file, there are 37 possible columns of information.

```
head(stormData, n = 2)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1 4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1 4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14   100 3   0          0
## 2         NA         0                       2   150 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1       15    25.0          K       0
## 2        0     2.5          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
```

The events in this database start in April 1950 and end in November 2011. A exploratory preview of the data suggested that the number of storm events did increase during this period. A subset from 1995-2011 was subsequently extracted for this analysis.

```
  stormData$year <- as.numeric(format(as.Date(stormData$BGN_DATE,
                                                format = "%m/%d/%Y %H:%M:%S"), "%Y"))

  #hist(stormData$year, breaks=30, col="lightblue", main="Storms by Year", xlab="Year")
  storm <- stormData[stormData$year >= 1995,]
```

Remove the above comment control to view the number of storms over the period 1950-2011 as a histogram

Methods

In this section, we check the number of personal **fatalities** and **injuries** that were caused by severe weather. The top 12 of the most severe types of weather events are reviewed.

```
fSortAssist <- function(fieldName, top=12, data=stormData)
  {
    index <- which(colnames(data) == fieldName)
    field <- aggregate(data[, index], by=list(data$EVTYPE), FUN="sum")
    names(field) <- c("EVTYPE", fieldName)
    field <- arrange(field, field[, 2], decreasing=TRUE)
    field <- head(field, n=top)
    field <- within(field, EVTYPE <- factor(x=EVTYPE, levels=field$EVTYPE))
    return(field)
  }

message("Stand By, processing data ...")
```

```
## Stand By, processing data ...
```

```
fatalities <- fSortAssist("FATALITIES", 12, data=storm)
injuries <- fSortAssist("INJURIES", 12, data=storm)
```

Impact on Economy

The **property** and **crop damage** data was converted into a monetary form according to the units described in the NOAA document, (NWSPD 10-16, NATIONAL WEATHER SERVICE INSTRUCTION 10-1605 (available%20at:%20http://www.nws.noaa.gov/directives/)). Both *PROPDMGEXP* (PROPerty DaMaGe EXPense) and *CROPDMGEXP* (CROP DaMaGe EXPense) columns record a cost multiplier for each event, where Hundred (H), Thousand (K), Million (M) and Billion (B) have their respective meaning. The following R-code snippet calculates the financial impact.

```
#cost x Hundred (H), Thousand (K), Million (M) and Billion (B).
fMonetaryAssist <- function(data=storm, fieldName, newFieldName)
  {
    t_Len <- dim(data)[2]
    index <- which(colnames(data) == fieldName)
    data[, index] <- as.character(data[, index])
    qNA <- !is.na(toupper(data[, index]))

    data[qNA & toupper(data[, index]) == "B", index] <- "9"   #billions
    data[qNA & toupper(data[, index]) == "M", index] <- "6"   #millions
    data[qNA & toupper(data[, index]) == "K", index] <- "3"   #thousands
    data[qNA & toupper(data[, index]) == "H", index] <- "2"   #hundreds
    data[qNA & toupper(data[, index]) == "" , index] <- "0"

    data[, index] <- as.numeric(data[, index])
    data[is.na(data[, index]), index] <- 0
    data <- cbind(data, data[, index - 1] * 10^data[, index])


    names(data)[t_Len + 1] <- newFieldName
    return(data)
  }

    storm <- fMonetaryAssist(storm, "PROPDMGEXP", "propertyDamage")
    storm <- fMonetaryAssist(storm, "CROPDMGEXP", "cropDamage")
    names(storm)
```

```
##  [1] "STATE__"         "BGN_DATE"        "BGN_TIME"        "TIME_ZONE"
##  [5] "COUNTY"          "COUNTYNAME"      "STATE"           "EVTYPE"
##  [9] "BGN_RANGE"       "BGN_AZI"         "BGN_LOCATI"      "END_DATE"
## [13] "END_TIME"        "COUNTY_END"      "COUNTYENDN"      "END_RANGE"
## [17] "END_AZI"         "END_LOCATI"      "LENGTH"          "WIDTH"
## [21] "F"               "MAG"             "FATALITIES"      "INJURIES"
## [25] "PROPDMG"         "PROPDMGEXP"      "CROPDMG"         "CROPDMGEXP"
## [29] "WFO"             "STATEOFFIC"      "ZONENAMES"       "LATITUDE"
## [33] "LONGITUDE"       "LATITUDE_E"      "LONGITUDE_"      "REMARKS"
## [37] "REFNUM"          "year"            "propertyDamage" "cropDamage"
```

```
    options(scipen=999)


    property <- NULL
    crop <- NULL
    property <- fSortAssist("propertyDamage", 12,  data=storm)
    crop <- fSortAssist("cropDamage", 12, data=storm)
```

Results

For the impact on public health, we have a table of severe weather events, merged around common event types. For the most severe, the number of people affected are enumerated. This allows viewing the physical (bodily) ‘damage’ side-by-side, assisting in comparision for probable events outcomes. Of note are the numbers for **Excessive Heat** and **Tornados**.

```
merge(fatalities, injuries)
```

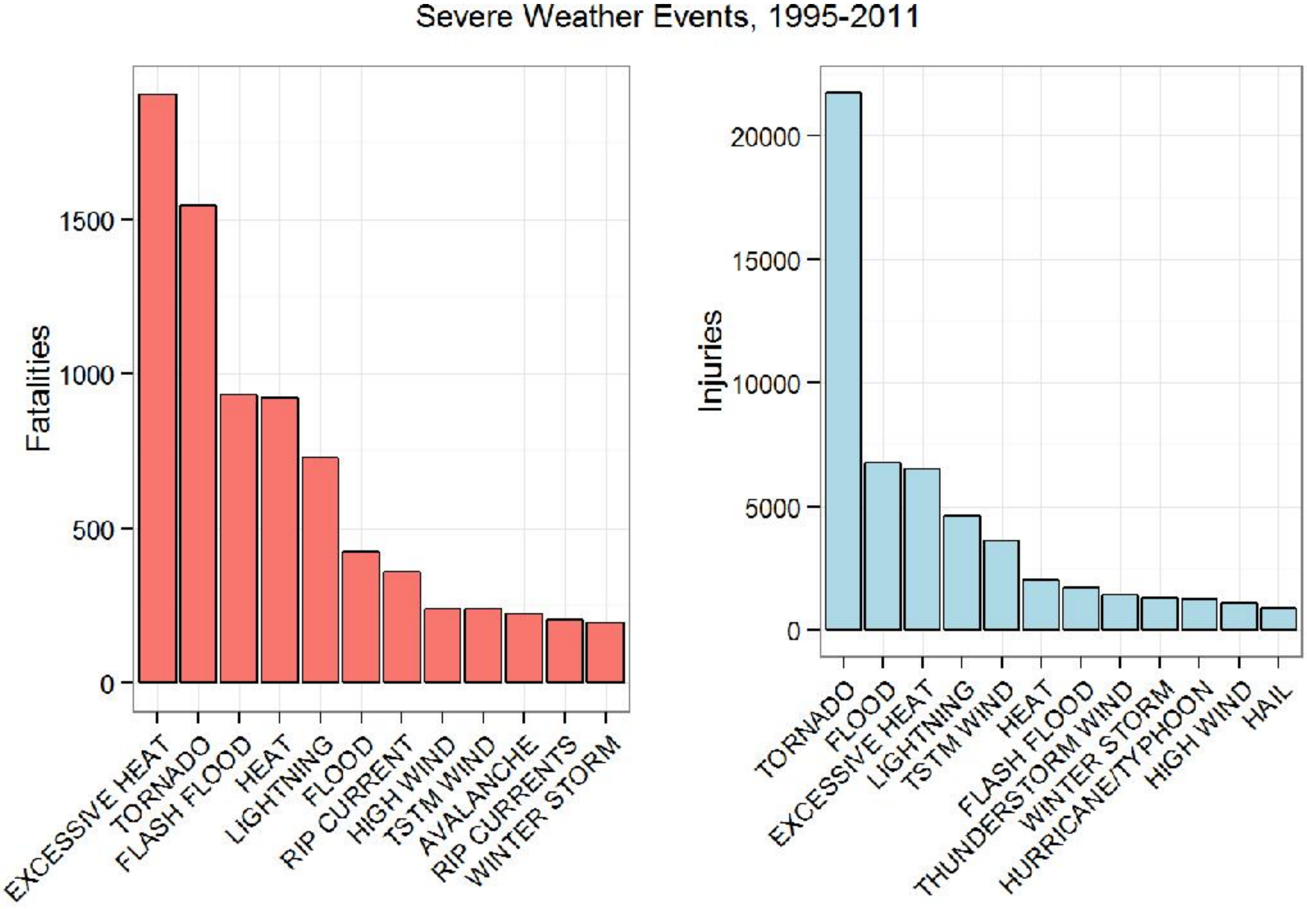
```
##           EVTYPE FATALITIES INJURIES
## 1 EXCESSIVE HEAT       1903     6525
## 2    FLASH FLOOD        934     1734
## 3          FLOOD        423     6769
## 4           HEAT        924     2030
## 5      HIGH WIND        241     1093
## 6      LIGHTNING        729     4631
## 7        TORNADO       1545    21765
## 8      TSTM WIND        241     3630
## 9   WINTER STORM        195     1298
```

Below is a set of graphs of fatalities and injuries from severe weather events.

```
#fatalities for 12 severe weather events summed across 1995-2011
pFatalities <- qplot(EVTYPE, data=fatalities, weight=FATALITIES, binwidth = 1) +
                    theme_bw() +
                    geom_bar(aes(fill="red"), color="black") +
                    scale_y_continuous("Fatalities") +
                    theme(axis.text.x = element_text(angle = 45, hjust = 1),
                          legend.position = "none") +
                    xlab("")

#injuries, with the same parameters
pInjuries <- qplot(EVTYPE, data=injuries, weight=INJURIES, binwidth = 1) +
                  theme_bw() +
                  geom_bar(fill="lightblue", color="black") +
                  scale_y_continuous("Injuries") +
                  theme(axis.text.x = element_text(angle = 45, hjust = 1),
                        legend.position = "none") +
                   xlab("")

#create a panel plot of the above two
grid.arrange(pFatalities, pInjuries, ncol = 2,
             main="Severe Weather Events, 1995-2011")
```



Find all weather events that are either hail or tornado by state. As there is a strong correlation between the occurance of damaging hail prior to a tornado, the data is scoured over the 1995-2011 time frame.

```
evt_dim <- 1:55
evt_names <- c("count1", "count2", "where", "StNum", "StName")
evt <- data.frame(evt_dim, evt_dim, evt_dim, evt_dim, evt_dim)
colnames(evt) <- evt_names
evt$StName <- NA


for (i in 1:55)
  {
    evt$count1[i] <- length(which(storm[,"EVTYPE"] == "HAIL" & storm[,"STATE__"] == i))
    evt$count2[i] <- length(which(storm[,"EVTYPE"] == "TORNADO" & storm[,"STATE__"] == i))
    evt$where[i]  <- i
  }

for (i in 1:nrow(storm))
  {
    idx <- storm[i,"STATE__"]
    if (is.na(evt$StName[idx]))
      {
        evt$StName[idx] <- as.character(storm[i, "STATE"])
        evt$StNum[idx]  <- i
      }
    if (idx == 55) break
  }
```

Scan the data for non-weather events, summary and Metro, and remove

```
unq <- unique(storm$EVTYPE)
unq[grep("^Summ|^SUMM|^Metr", unq)] = NA   # remove all 'summary|SUMMARY|Metro', keep others
unq <- unq[!is.na(unq[])]
head(unq, 20)
```

```
##  [1] "FREEZING RAIN"                  "SNOW"
##  [3] "SNOW/ICE"                       "HURRICANE OPAL/HIGH WINDS"
##  [5] "HAIL"                           "THUNDERSTORM WINDS"
##  [7] "RECORD COLD"                    "HURRICANE ERIN"
##  [9] "HURRICANE OPAL"                 "DENSE FOG"
## [11] "RIP CURRENT"                    "TORNADO"
## [13] "THUNDERSTORM WINS"              "LIGHTNING"
## [15] "FLASH FLOOD"                    "FLASH FLOODING"
## [17] "HIGH WINDS"                     "TORNADO F0"
## [19] "THUNDERSTORM WINDS LIGHTNING" "FUNNEL CLOUD"
```

Find all the Hail and Tornado events

```
Hail_Events <- unq[grep("^HAIL", unique(storm$EVTYPE))]
Tornado_Events <- unq[grep("^TORN", unique(storm$EVTYPE))]
```

Show a few types of these events

```
Hail_Events
```

```
##  [1] "HAIL"                    "HAIL STORM"
##  [3] "HAIL 75"                 "HAIL 80"
##  [5] "HAIL 0.75"               "HAIL 1.00"
##  [7] "HAIL 1.75"               "HAIL 88"
##  [9] "HAIL 175"                "HAIL 100"
## [11] "HAIL 150"                "HAIL 075"
## [13] "HAIL 125"                "HAIL 200"
## [15] "HAIL FLOODING"           "HAIL 088"
## [17] "HAIL 275"                "HAIL 450"
## [19] "HAILSTORM"               "HAILSTORMS"
## [21] "HAIL DAMAGE"             "URBAN/SML STREAM FLDG"
```

```
Tornado_Events
```

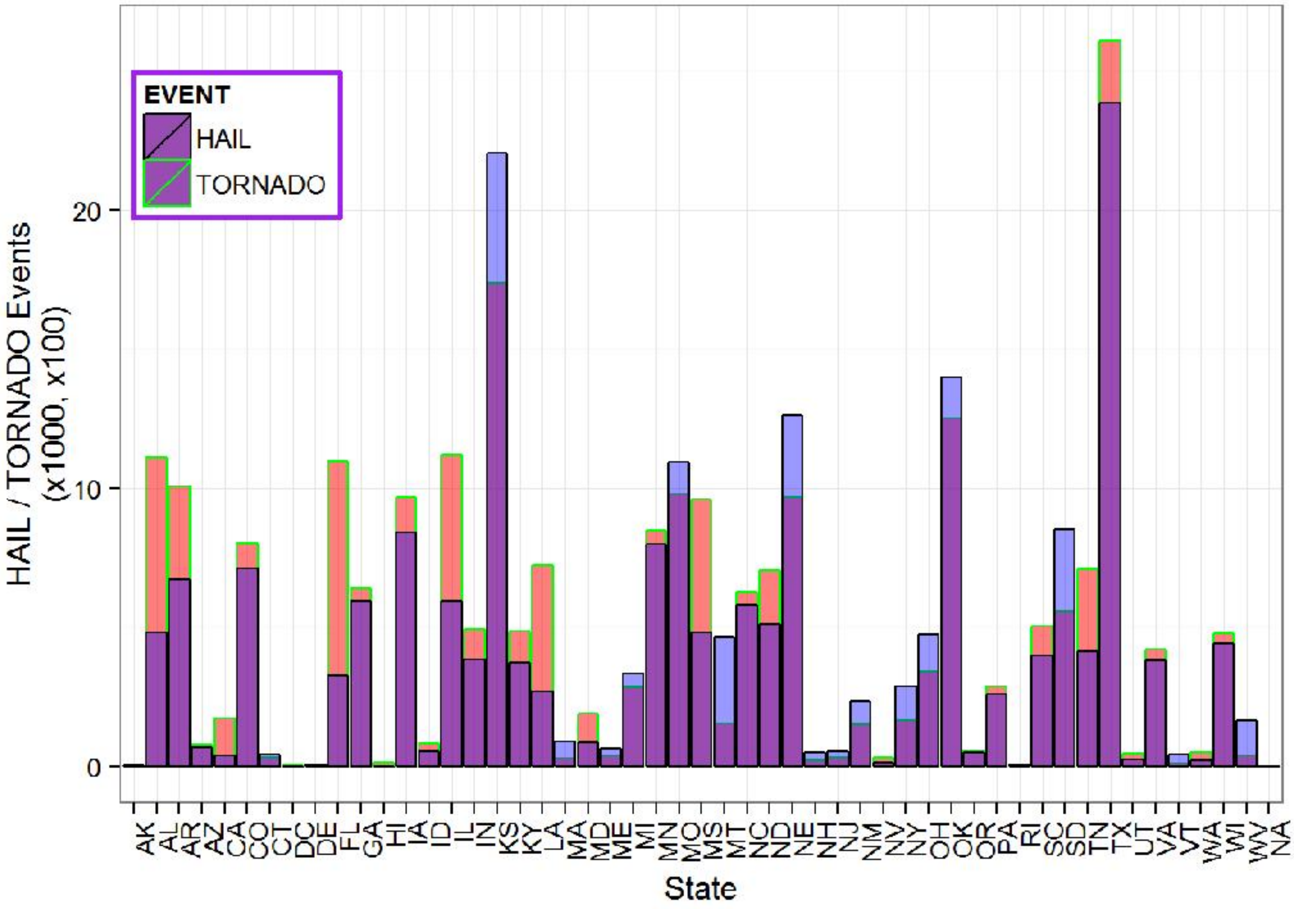
```
## [1] "TORNADO"             "TORNADO F0"          "TORNADOS"
## [4] "TORNADO F3"          "TORNADO F1"          "TORNADO/WATERSPOUT"
## [7] "TORNADO F2"          NA
```

This plot depicts the total hail and tornado events across the country during the study period. First, note the legend in the upper left. It shows the outline of each event, either black of green outlined. Hail is a lightblue bar with the black border and the tornado is a redish bar with the green border. They are semi-transparent, so tht neither can block the other and it can be seen which is more revalent for a particular state during the study period. Of interest is the hail vs tornado coupling, which is important for economic and insurance forcasting, and avoidance for individuals during such events.

```
hail = "blue"
tornado = "red"

t_HailTornadoByState<-ggplot() + theme_bw() +
                  geom_bar(data=evt, aes(x=StName, y=count2/100, color=tornado),
                           fill=tornado, stat="identity", alpha=0.5) + #tornado
                  geom_bar(data=evt, aes(x=StName, y=count1/1000, color=hail),
                           fill=hail, stat="identity", alpha=0.4) + #hail
                  scale_y_continuous("HAIL / TORNADO Events\n(x1000, x100)") +
                  theme(axis.text.x = element_text(angle = 90, hjust = 1),
                        legend.position=c(0.1, 0.825)) +
                  scale_color_manual(name = "EVENT", values = c("black", "green"), labels=c("HAIL",
 "TORNADO")) +
                  theme(legend.background = element_rect(color = 'purple', fill = 'white', size = 1
)) +
                  scale_fill_manual(values=c(hail="blue", tornado="red"),
                                    guide="none") +
                  xlab("State")

print(t_HailTornadoByState)
```



Based on the above histograms, we find that **excessive heat** and **tornados** cause most fatalities; tornados caused most injuries in the United States from 1995 to 2011. This seems counter intuitive, as one might consider hurricanes the most damaging and dangerous. However, tornado appearances can have very little warning. This happens especially during the night, when most individuals are not quite awake when they hear a siren or the roar of the high velocity winds.

```
printMoney <- function(x)
  {
    format(x, digits=10, nsmall=2, decimal.mark=".", big.mark=",")
  }

t_property <- property
t_crop<- crop

#change column names to '$'name
npd <- names(t_property)[2]; new_npd <- paste("$", npd, sep=""); names(t_property)[2] <- new_npd
ncd <- names(t_crop)[2];     new_ncd <- paste("$", ncd, sep=""); names(t_crop)[2] <- new_ncd

#select weather events common to property and crop damage
mpc <- merge(t_property, t_crop)

#calc ratio between property and crops
mpc[,4] <- round(mpc[,2]/mpc[,3], 2)
colnames(mpc)[4] <- "ratio"

#format property and crop numbers to $$ amounts with '.,'
mpc[,2] <- printMoney(mpc[,2])
mpc[,3] <- printMoney(mpc[,3])
```

For the economic impact, see the table below. The list displays the common damage events, e.g. FLOOD, between property and crops. There is a large difference in damage estimates, as noted by the ‘ratio’ column in the data table.

```
mpc
```

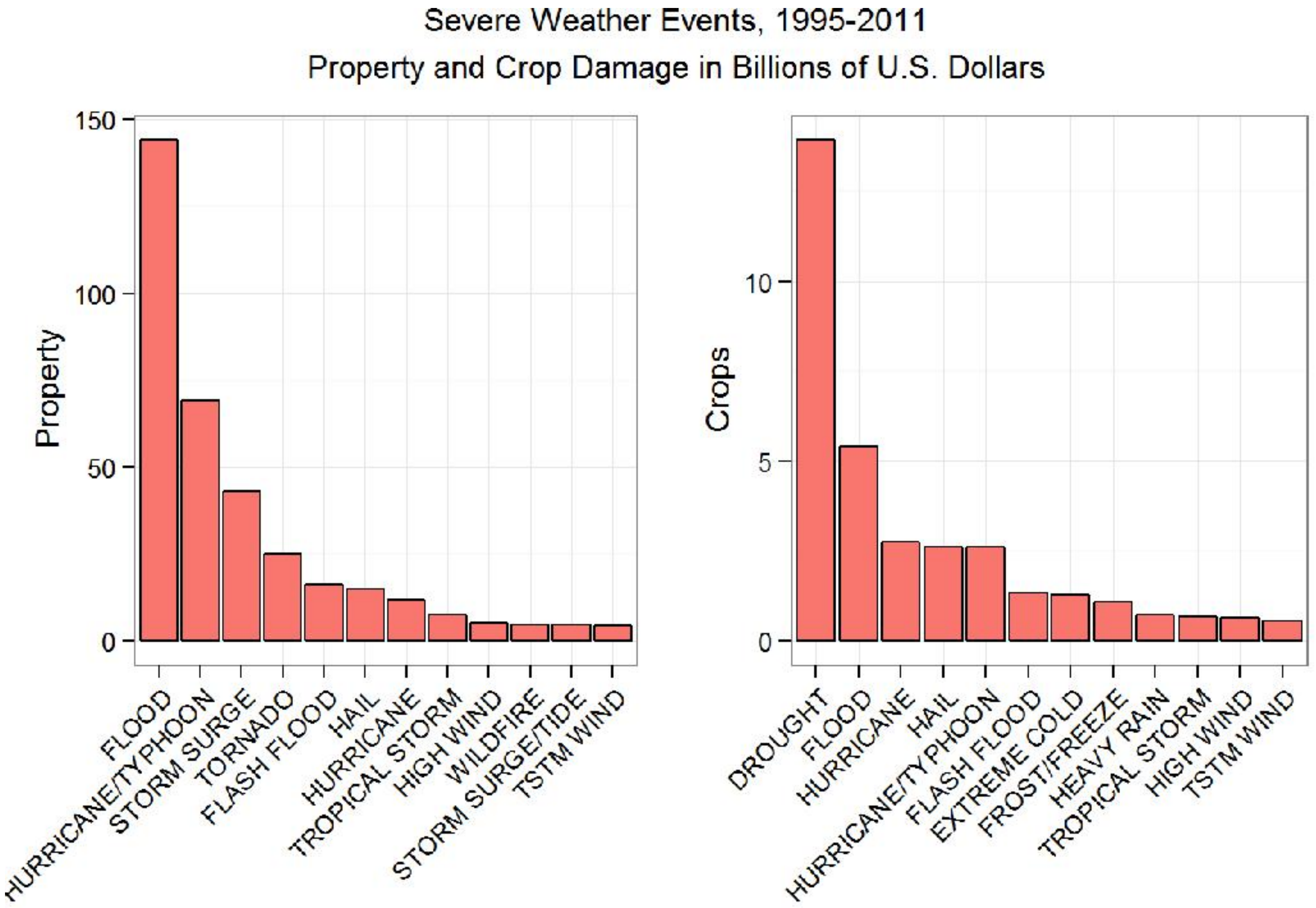
```
##              EVTYPE    $propertyDamage       $cropDamage ratio
## 1       FLASH FLOOD  16,047,794,571.00 1,343,915,000.00 11.94
## 2             FLOOD 144,022,037,057.00 5,422,810,400.00 26.56
## 3              HAIL  15,048,722,102.70 2,614,127,070.00  5.76
## 4         HIGH WIND   5,259,785,375.00   633,561,300.00  8.30
## 5         HURRICANE  11,812,819,010.00 2,741,410,000.00  4.31
## 6 HURRICANE/TYPHOON  69,305,840,000.00 2,607,872,800.00 26.58
## 7    TROPICAL STORM   7,653,335,550.00   677,836,000.00 11.29
## 8         TSTM WIND   4,482,361,440.00   553,947,350.00  8.09
```

The following is a pair of graphs of total property and crop damage affected by severe weather events.As mentioned previously, property damage can far exceed crop damage, and this fact is clearly illustrated in the graphs. Note the property related monetary outcome of FLOOD and HURRICANE related events. DROUGHT is the strongest indicator for crops, as lack of water is the strongest detriment to plant growth (in the context discussed).

```
propertyDamage <- qplot(EVTYPE, data=property, weight=propertyDamage/1e+09, binwidth=1) +
                        theme_bw() + theme(legend.position = "none") +
                        geom_bar(aes(fill="red"), color="black") +
                        theme(axis.text.x=element_text(angle=45, hjust=1)) +
                        ylab(expression(paste("Property", sep=""))) +
                        xlab("")

cropDamage <- qplot(EVTYPE, data=crop, weight=cropDamage/1e+09, binwidth=1) +
                    theme_bw() + theme(legend.position = "none") +
                    geom_bar(aes(fill="lightblue"), color="black") +
                    theme(axis.text.x=element_text(angle=45, hjust=1)) +
                    ylab(expression(paste("Crops", sep=""))) +
                    xlab("")

grid.arrange(propertyDamage, cropDamage, ncol = 2,
             main=("Severe Weather Events, 1995-2011\nProperty and Crop Damage in Billions of U.S.
Dollars"))
```



Conclusion

The emphasis of this report has been to analyze and display important information gleaned from the NOAA Storm database, extracting the 1995 to 2011 subset. The R statistical programming language along with the RStudio environment, were utilized to perform both tasks, as well as acquiring the database. During the exploratory data analysis, several facets were noticed, damage from various storm events, property vs. crop economics, coupling of storm activities, and fatalities and injuries from weather events.

For injuries and fatalities, the two most distructive weather forces are EXCESSIVE HEAT, such as drought, and TORNADOs. High wind events are ~58% of the total, with ~$106 billion dollars of damage to property and crops. Flooding, however, causes the most damage, financially, to property and crops. See the above table.

Considering the above ‘Severe Weather Events’ histograms, they indicate water related events cause the most property and crop damage while heat and tornados are difficult to withstand for personel.

Of particular interest was the coupling of hail with tornados, which is implied by the HAIL/TORNADO Events by State histogram. There were four states, Kansas, Missouri, Nebraska and Oklahoma, that had a higher incidence of hail events to tornados. These states are in the middle of ‘Tornado Alley’, the worst place for tornado occurances in the Unied States. The ‘disparity’ between incidences of hail to tornados is explained because, a thunderstorm can produce hail due to the recycling up-draft present in the storm’s center, while the same storm does not necessarily produce a tornado, as this event requires a lateral to vertical wind field shift, which is not always present. Texas had the most tornado and hail events over the study period, and is also in the danger zone.