An Examination of the Health and Economic Impact of Weather

Synopsis

in the following analysis, we will examine a collection of weather event data from the US between the years of 1950 and 2011. Through this analysis we have attempted to address:

- What type of weather has the biggest health impact?
- What type of weather has the biggest economic impact?

Our findings show that Tornadoes appear to have the highest impact on population health, both in terms of injuries and fatalities. However, flooding and hurricanes appear to have the largest economic impact.

Data Processing

The data used in this analysis is located at: link

We'll download, unarchive, and read in the data to a variable called `stormData`:

```
archiveFile <- "StormData.csv.bz2"
download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", archiveFile)
rawStormData <- read.csv(archiveFile)
```

R makes some intelligent guesses about data types when reading in data, but we also need to do some housekeeping. Function definitions for `cleanStr`, `mapEVTYPE`, `getExp` are shown in Appendix 1.

```
stormData <- rawStormData

# Convert BGN_DATE factor -> datetime
stormData$BGN_DATE <- date(mdy_hms(stormData$BGN_DATE))

# Convert END_DATE factor -> datetime
stormData$END_DATE <- date(mdy_hms(stormData$END_DATE))
# Fix missing END_DATE
# ASSUMPTIONS:
# Where END_DATE is not specified in the database, I make the assumption
# that the event does not last longer than 24 hours. In essence, I assign
# END_DATE = BGN_DATE
stormData$END_DATE[is.na(stormData$END_DATE)] <-
  stormData$BGN_DATE[is.na(stormData$END_DATE)]

# Clean EVTYPE
stormData$EVTYPE <- unlist(lapply(stormData$EVTYPE, cleanStr))
stormData$EVTYPE <- factor(mapEVTYPE(stormData$EVTYPE))

# Clean PROPDMG
stormData$PROPDMGEXP <- unlist(lapply(stormData$PROPDMGEXP, cleanStr))
stormData$PROPDMGEXPN <- getExp(stormData$PROPDMGEXP)
stormData$PROPDMG <- stormData$PROPDMG * stormData$PROPDMGEXPN
```

```
# Clean CROPDMG
stormData$CROPDMGEXP <- unlist(lapply(stormData$CROPDMGEXP, cleanStr))
stormData$CROPDMGEXPN <- getExp(stormData$CROPDMGEXP)
stormData$CROPDMG <- stormData$CROPDMG * stormData$CROPDMGEXPN

# Create TOTALDMG
stormData$TOTALDMG <- stormData$CROPDMG + stormData$PROPDMG

# Remove unused columns
usedCols = c("BGN_DATE", "TIME_ZONE", "STATE", "EVTYPE", "END_DATE",
             "LENGTH", "WIDTH", "F", "MAG", "FATALITIES", "INJURIES",
             "PROPDMG", "CROPDMG", "TOTALDMG")
newStormData <- stormData[, usedCols]
```

The code shown above does the following:

- Dates (`BGN_DATE`, `END_DATE`) are converted from factors into datetime instances
  - `NA` values are handled in `END_DATE` by assuming the event lasts less than 24 hours
- Event types are cleaned up into predefined categories
  - Not all events are cleaned, but the majority fit into fixed categories
  - Category names are as defined in the database documentation here
- Economic costs are cleaned and simplified
  - `CROPDMGEXP` and `PROPDMGEXP` are converted from character values into powers of ten
  - `CROPDMG` and `PROPDMG` are scaled by the appropriate powers of ten above
- `TOTALDMG` with the sum of `CROPDMG` and `PROPDMG` is created
- A simplified version of the database is created with a subset of columns to be analyzed

Some columns were removed due to redundancy (e.g. `CROPDMGEXP`), while others were removed due to limited usefulness (e.g. `REMARKS`)

Results

By quantitative analysis of the NOAA dataset, we will attempt evaluate the following:

- Which types of weather events are most harmful to the health of the US population?
- Which types of events have the greates economic consequences?

Event Health Impact

The NOAA dataset contains two different metrics by which we can measure impact to population health: fatalities and injuries. While fatalities are certainly a devastating impact of some weather events, they do not necessarily tell the whole story. The additional healthcare expense of treating injuries of people affected by weather events can translate to a heavier economic burden in a local area. For this reason, we evaluate weather event severity based on both metrics.

There are many metrics we could use to evaluate the relative fatality and injury counts of weather events. For our analysis, we chose to use maximum fatalities/injuries from a single event (severity) and total fatalities/injuries between 1950-2011 (frequency). These two summaries of the data may give local officials the best idea where investments can be made to safeguard the population.

Let's look at the top five weather events for these four metrics:

Figure 1: Top Ten Weather Event Types by Fatalities

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

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

```
library(pander)

topFatalities <- stormData %>%
  group_by(EVTYPE) %>%
  summarize(total_fatalities = sum(FATALITIES, na.rm=TRUE),
            max_fatalities = max(FATALITIES, na.rm=TRUE)) %>%
  filter(total_fatalities>0) %>%
  arrange(desc(total_fatalities), desc(max_fatalities)) %>%
  mutate("Rank"=row_number()) %>%
  top_n(n = 10, wt = total_fatalities) %>%
  rename("Fatalities due to Single Event" = max_fatalities, "Total Fatalities 1950-2011" = total_fatali
pander(topFatalities)
```

Event Type	Total Fatalities 1950-2011	Fatalities due to Single Event	Rank
TORNADO	5633	158	1
EXCESS HEAT	3132	583	2
FLASH FLOOD	1035	20	3
LIGHTNING	817	5	4
THUNDERSTORM	726	11	5
FLOOD	490	15	6
STRONG WIND	419	8	7
RIP CURRENT	368	6	8
WINTER STORM	279	10	9
WIND CHILL	237	10	10

Figure 2: Top Ten Weather Event Types by Injuries

```
topInjuries <- stormData %>%
  group_by(EVTYPE) %>%
  summarize(total_injuries = sum(INJURIES, na.rm=TRUE),
            max_injuries = max(INJURIES, na.rm=TRUE)) %>%
  filter(total_injuries>0) %>%
  arrange(desc(total_injuries), desc(max_injuries)) %>%
  mutate("Rank"=row_number()) %>%
  top_n(n = 10, wt = total_injuries) %>%
  rename("Injuries due to Single Event" = max_injuries, "Total Injuries 1950-2011" = total_injuries, "E
pander(topInjuries)
```

Event Type	Total Injuries 1950-2011	Injuries due to Single Event	Rank
TORNADO	91364	1700	1
THUNDERSTORM	9449	70	2
EXCESS HEAT	9209	519	3
FLOOD	6802	800	4
LIGHTNING	5231	51	5
ICE STORM	2154	1568	6
WINTER STORM	1968	165	7
STRONG WIND	1830	89	8
FLASH FLOOD	1802	150	9
WILDFIRE	1606	150	10

In summary, we can see that by both metrics, Tornadoes appear to be most impactful in terms of total numbers over a sixty year period. It is interesting to note, however, that the highest number of fatalities due to a single event is attributed to excess heat.

Event Economic Impact

We can examine the economic impact of weather events using a similar strategy.

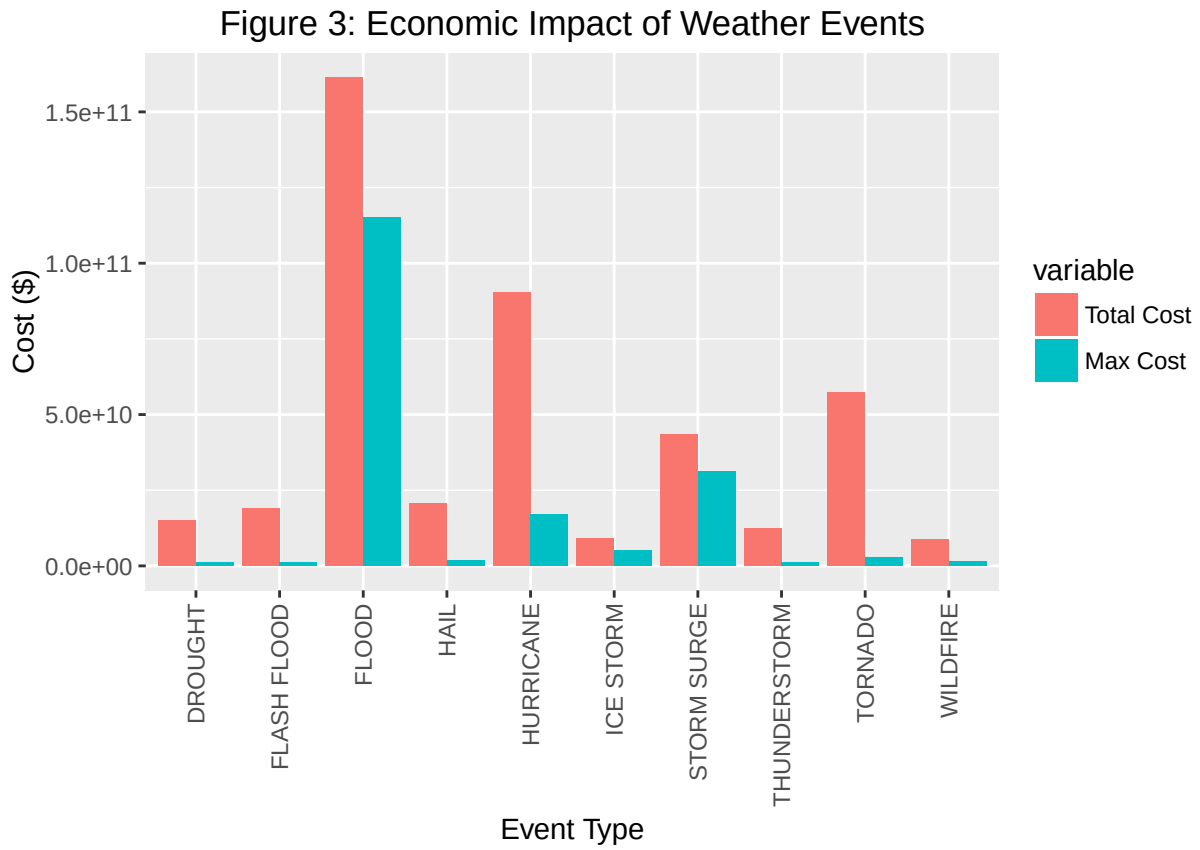
```
library(ggplot2)
library(dplyr)
library(reshape2)

topEconEvents <- stormData %>%
  group_by(EVTYPE) %>%
  summarize(total_cost = sum(TOTALDMG, na.rm=TRUE),
            max_cost = max(TOTALDMG, na.rm=TRUE)) %>%
  filter(total_cost>0) %>%
  top_n(n = 10, wt = total_cost) %>%
  rename("Total Cost" = total_cost, "Max Cost" = max_cost)

topEconEvents.long <- melt(topEconEvents, id.vars="EVTYPE")
```

```
topEconEvents.long$EVTYPE <- factor(topEconEvents.long$EVTYPE)
topEconEvents.long$variable <- factor(topEconEvents.long$variable)

plt <- ggplot(topEconEvents.long, aes(x=EVTYPE, y=value, fill=variable)) +
  geom_bar(stat="identity", position="dodge") +
  scale_color_discrete("Metric") +
  xlab("Event Type") +
  ylab("Cost ($)") +
  ggtitle("Figure 3: Economic Impact of Weather Events") +
  theme(axis.text.x=element_text(angle=90, hjust=1, vjust=0.5))
plt
```



There are a few observations that can be made from this plot. First, we note that flooding appears to have, by far, the most significant economic impact. However, we also note that a significant portion of this total cost is attributed to a single event. Although hurricanes fall far short of the total cost, the maximum cost of a single event is far lower. This would suggest hurricanes occur with much higher frequency and may add up to a larger number over time.

Appendix 1

```
library(lubridate)

# Utility function for cleaning strings
#
```

```
cleanStr <- function(s) {
  s <- gsub("^\\s+|\\s+$", "", s)
  s <- gsub("[[:punct:]]", " ", s)
  toupper(s)
}


# Utility function for cleaning EVTYPE
#
# ASSUMPTIONS:
# The data input into this column is highly irregular and does not
# conform to a particular set of valid entries. I have attempted to
# standardize the names as much as possible. There are still a large
# number of entries with unique values that have not been addressed.
# In addition, it is possible that some events have multiple types
# in the description. I make no assumption about "importance" when
# assigning the single category to an event. Names are filtered
# roughly alphabetically using `grep`, and secondary event types
# are most likely removed.
#
mapEVTYPE <- function(v) {
  v[grep("ASTRO.+LOW", v)] <- "ASTRONOMICAL LOW TIDE"
  v[grep("ASTRO.+HIGH", v)] <- "ASTRONOMICAL HIGH TIDE"
  v[grep("AVAL", v)] <- "AVALANCHE"
  v[grep("BLIZ", v)] <- "BLIZZARD"
  v[grep("COAST.+FLOOD", v)] <- "COASTAL FLOOD"
  v[grep("CHILL", v)] <- "WIND CHILL"
  v[grep("DENSE.+FOG", v)] <- "DENSE FOG"
  v[grep("SMOKE", v)] <- "SMOKE"
  v[grep("DROUG", v)] <- "DROUGHT"
  v[grep("DEV[IE]L", v)] <- "DUST DEVIL"
  v[grep("DUST *ST", v)] <- "DUST STORM"
  v[grep("SAHARA", v)] <- "DUST STORM"
  v[grep("EXCESS.+HEAT", v)] <- "EXCESS HEAT"
  v[grep("EXTR.+COLD", v)] <- "EXTREME COLD"
  v[grep("FLASH", v)] <- "FLASH FLOOD"
  v[setdiff(grep("FLOOD", v), grep("FLASH", v))] <- "FLOOD"
  v[grep("FRE.+FOG", v)] <- "FREEZING FOG"
  v[grep("FROS", v)] <- "FROST/FREEZE"
  v[grep("FUN", v)] <- "FUNNEL CLOUD"
  v[grep("HAIL", v)] <- "HAIL"
  v[grep("HEAT", v)] <- "EXCESS HEAT"
  v[grep("HEAV.+(RAIN|PREC)", v)] <- "HEAVY RAIN"
  v[grep("HEAV.+SNOW", v)] <- "HEAVY SNOW"
  v[grep("SURF", v)] <- "HEAVY SURF"
  v[grep("HURR", v)] <- "HURRICANE"
  v[grep("ICE", v)] <- "ICE STORM"
  v[grep("SLEE", v)] <- "SLEET"
  v[grep("THUNDER", v)] <- "THUNDERSTORM"
  v[grep("TSTM", v)] <- "THUNDERSTORM"
  v[grep("WATERSPOUT", v)] <- "WATERSPOUT"
  v[grep("TORN", v)] <- "TORNADO"
  v[grep("DEPR", v)] <- "TROPICAL DEPRESSION"
```

```
  v[grep("TROP.+STORM", v)] <- "TROPICAL STORM"
  v[grep("TSU", v)] <- "TSUNAMI"
  v[grep("VOL.+ASH", v)] <- "VOLCANIC ASH"
  v[grep("WILD", v)] <- "WILDFIRE"
  v[grep("WINT", v)] <- "WINTER STORM"
  v[grep("LIGHTN", v)] <- "LIGHTNING"
  v[grep("HIGH.+WIND", v)] <- "STRONG WIND"
  v[grep("STRONG", v)] <- "STRONG WIND"
  v[grep("SUMMARY", v)] <- NA
  return(v)
}


# Function for extracting exponent for PROPDMG and CROPDMG
#
# ASSUMPTIONS:
# There are both numeric and character values in the
# PROPDMGEXP and CROPDMGEXP variables. For numeric values,
# I have assumed that this is a power of ten multiplying
# PROPDMG or CROPDMG. For character values, I have assumed
# this variable refers to a particular power of ten
# (e.g. "M" -> Million -> $1,000,000 x PROPDMG).
# Where the value does not fit either of these criteria,
# I ignore the value of this variable (i.e, exponent = 1)
#
getExp <- function(v) {
  v[grep("0", v)] <- "1"
  v[grep("1", v)] <- "10"
  v[grep("[2H]", v)] <- "100"
  v[grep("[3K]", v)] <- "1000"
  v[grep("4", v)] <- "10000"
  v[grep("5", v)] <- "100000"
  v[grep("[6M]", v)] <- "1000000"
  v[grep("7", v)] <- "10000000"
  v[grep("8", v)] <- "100000000"
  v[grep("[9B]", v)] <- "1000000000"
  v <- as.numeric(v)
  v[is.na(v)] <- 1
  return(v)
}
```