Reproducible Research Project 2, Storm Data Analysis

Synopsis

This paper makes an attempt to explore the NOAA Storm Database and answer some basic questions about severe weather events.It uses data from U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. It employs an exploratory analysis to find which types of events are most harmful with respect to population health and which are for economic. We find that tornado is most harmful for population health considering both injuries and fatality. For economic losses in general, flood stands first, followed by hurricane/typhoon. For crop loss, it is drought and for housing loss, it is tornado again.

Data Processing

We have first downloaded the file from website using download.file() command and then extracted it using read.csv() function. Then we loaded all the necessary libraries for this analysis. Then we manipulate dataset so as to address our question.

```
url = "http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
download.file(url, dest = "storm.bz2")
data = read.csv(bzfile("storm.bz2"))
library(ggplot2)
library(plyr)
#setwd("C:/Data Science Specialization/Reproducible Research/Quiz/Peer Assessments/2/")
```

At first, to address the first question of weather events that are most harmful to population, we look at total number of injuries and fatalities by different weather events.

```
injuries = ddply(data, .(EVTYPE), summarize, sum.injuries = sum(INJURIES,na.rm=TRUE))
injuries = injuries[order(injuries$sum.injuries, decreasing = TRUE), ]
```

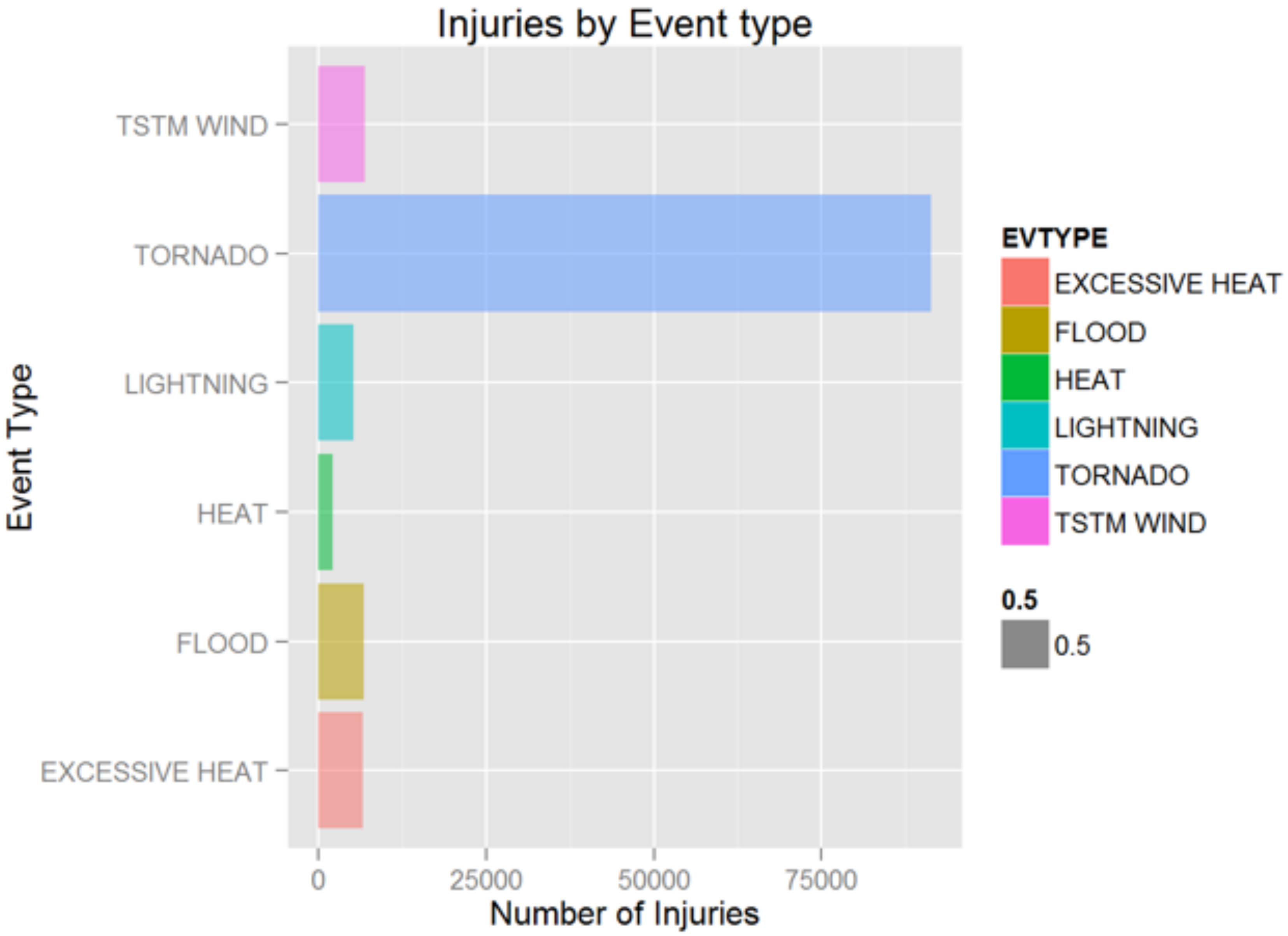
Now let's look at the 5 most harmful weather events.

```
head(injuries, 5)
```

```
##             EVTYPE sum.injuries
## 834        TORNADO        91346
## 856      TSTM WIND         6957
## 170          FLOOD         6789
## 130 EXCESSIVE HEAT         6525
## 464      LIGHTNING         5230
```

We see that tornado is the most harmful event with injuries of more than 91 thousands. This can be represented in the below figure:

```
ggplot(injuries[1:6, ], aes(EVTYPE, sum.injuries, fill = EVTYPE,alpha=0.5)) + geom_bar(stat = "identity") +
  xlab("Event Type") + ylab("Number of Injuries") + ggtitle("Injuries by Event type") + coord_flip()
```



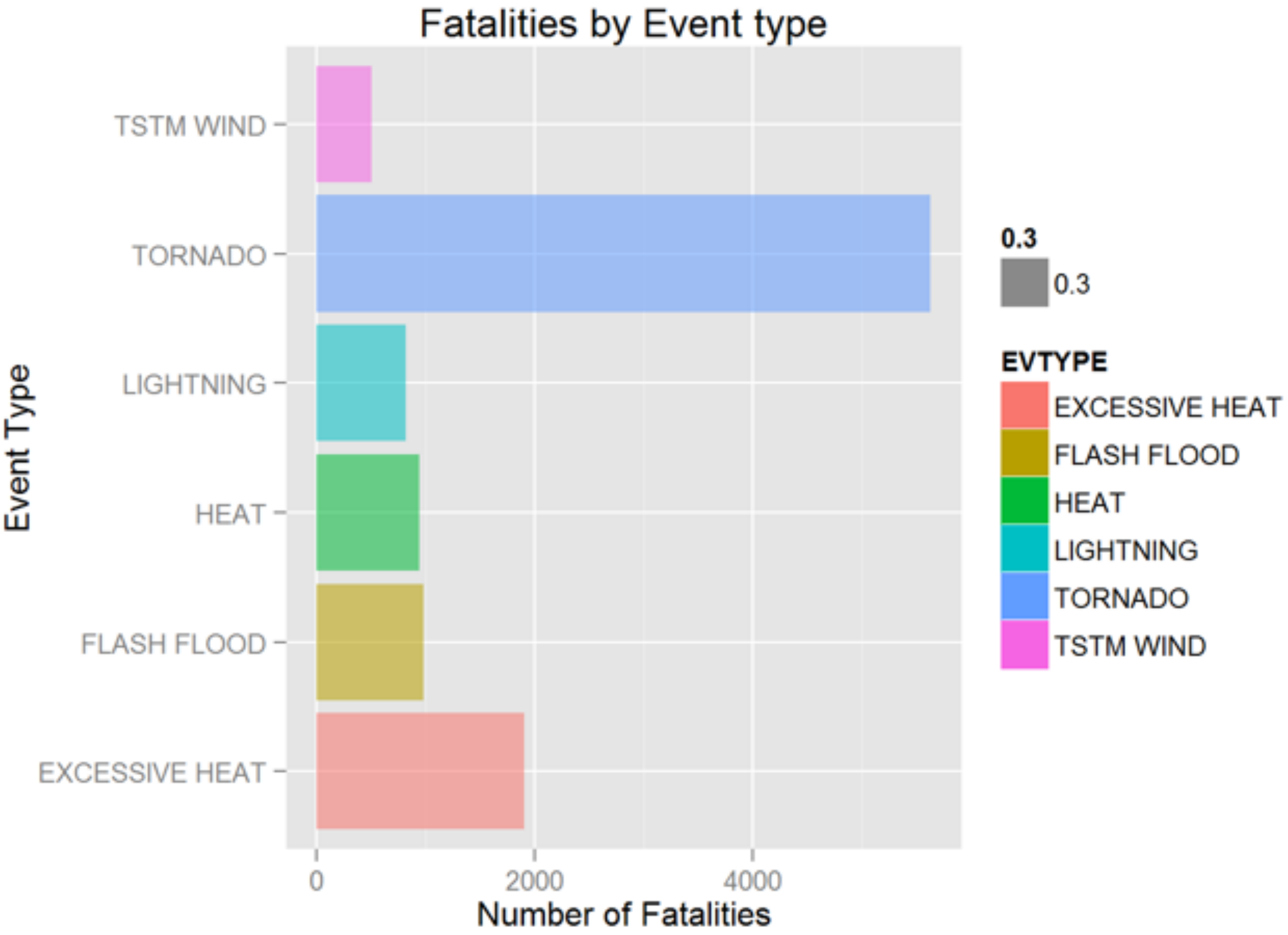
Now we will check for fatalities and look at the 5 most fatalities events.

```
fatalities = ddply(data, .(EVTYPE), summarize, sum = sum(FATALITIES))
fatalities = fatalities[order(fatalities$sum, decreasing = TRUE), ]
head(fatalities, 5)
```

```
##             EVTYPE  sum
## 834        TORNADO 5633
## 130 EXCESSIVE HEAT 1903
## 153    FLASH FLOOD  978
## 275           HEAT  937
## 464      LIGHTNING  816
```

We see that it is tornado again with fatalities of more than 5 thousands followed by excessive heat causing close to 2 thousands fatalities. We again provide a figure below to give a clear picture in a succint manner.

```
ggplot(fatalities[1:6, ], aes(EVTYPE, sum, fill=EVTYPE,alpha=0.3)) + geom_bar(stat = "ident
ity") +
  xlab("Event Type") + ylab("Number of Fatalities") + ggtitle("Fatalities by Event type") +
coord_flip()
```



To address the second question of economic consequences, we will investigate property damagea followed by crop damage and then total damage. Let's focus on property damage first. We start by looking at various exponents for PROPDMGEXP.

```
unique(data$PROPDMGEXP)
```

```
##  [1] K M   B m + 0 5 6 ? 4 2 3 h 7 H - 1 8
## Levels:  - ? + 0 1 2 3 4 5 6 7 8 B h H K m M
```

As some have lower character, we convert them to upper character. Also we
replace symbols other than character of numeric values to 0.

```
data$PROPDMGEXP <- toupper(data$PROPDMGEXP)
data$PROPDMGEXP[data$PROPDMGEXP %in% c("", "+", "-", "?")] = "0"
```

As PROPDMGEXP stands for the power of 10, we convert ‘B’ standing for billions to 9, ‘M’ standing for millions to 6, ‘K’ standing for thousands to 3 and ‘H’ for hundreds to 2.

```
data$PROPDMGEXP[data$PROPDMGEXP %in% c("B")] = "9"
data$PROPDMGEXP[data$PROPDMGEXP %in% c("M")] = "6"
data$PROPDMGEXP[data$PROPDMGEXP %in% c("K")] = "3"
data$PROPDMGEXP[data$PROPDMGEXP %in% c("H")] = "2"
```

Now we get the full property damage by converting PROPDMGEXP to numeric values and calculating total damage by multiplying the damage by the corresponding exponent.

```
data$PROPDMGEXP <- 10^(as.numeric(data$PROPDMGEXP))
damage.property = data$PROPDMG * data$PROPDMGEXP
data=as.data.frame(cbind(data,damage.property))
```

Now we make a new dataset of property damage arranged according to events type and look at the first 6 major events in terms of economic loss.

```
Damage.property = ddply(data, .(EVTYPE), summarize, damage.property = sum(damage.property,
na.rm = TRUE))
# Sort the Damage dataset
Damage.property = Damage.property[order(Damage.property$damage.property, decreasing = T), ]
# Show the first 6 most damaging types
head(Damage.property)
```

```
##                EVTYPE damage.property
## 170             FLOOD       1.447e+11
## 411 HURRICANE/TYPHOON       6.931e+10
## 834           TORNADO       5.695e+10
## 670       STORM SURGE       4.332e+10
## 153       FLASH FLOOD       1.682e+10
## 244              HAIL       1.574e+10
```

We see that Flood is the major damaging event for housing in terms of economic loss with a total amount of more than 144 billion. This is followed by hurricane/typhoon and tornado.

Now we will look at which event is most devastating economically for crops. As with the economic computation, we take the similar steps and look at the most damaging event for crops in terms of economic loss.

Let’s have a look at various exponents for CROPDMGEXP.

```
unique(data$CROPDMGEXP)
```

```
## [1]   M K m B ? 0 k 2
## Levels:  ? 0 2 B k K m M
```

As two levels have lower characters, we convert them to upper character. Also we replace symbols other than character of numeric values to 0.

```
data$CROPDMGEXP <- toupper(data$CROPDMGEXP)
data$CROPDMGEXP[data$CROPDMGEXP %in% c("", "?")] = "0"
```

As PROPDMGEXP stands for the power of 10, we convert ‘B’ standing for billions to 9, ‘M’ standing for millions to 6, ‘K’ standing for thousands to 3 and ‘H’ for hundreds to 2.

```
data$CROPDMGEXP[data$CROPDMGEXP %in% c("B")] = "9"
data$CROPDMGEXP[data$CROPDMGEXP %in% c("M")] = "6"
data$CROPDMGEXP[data$CROPDMGEXP %in% c("K")] = "3"
data$CROPDMGEXP[data$CROPDMGEXP %in% c("H")] = "2"
```

Now we get the full crop damage by converting PROPDMGEXP to numeric values and calculating total damage by multiplying the damage by the corresponding exponent.

```
data$CROPDMGEXP <- 10^(as.numeric(data$CROPDMGEXP))
damage.crop = data$CROPDMG * data$CROPDMGEXP
data=as.data.frame(cbind(data,damage.crop))
```

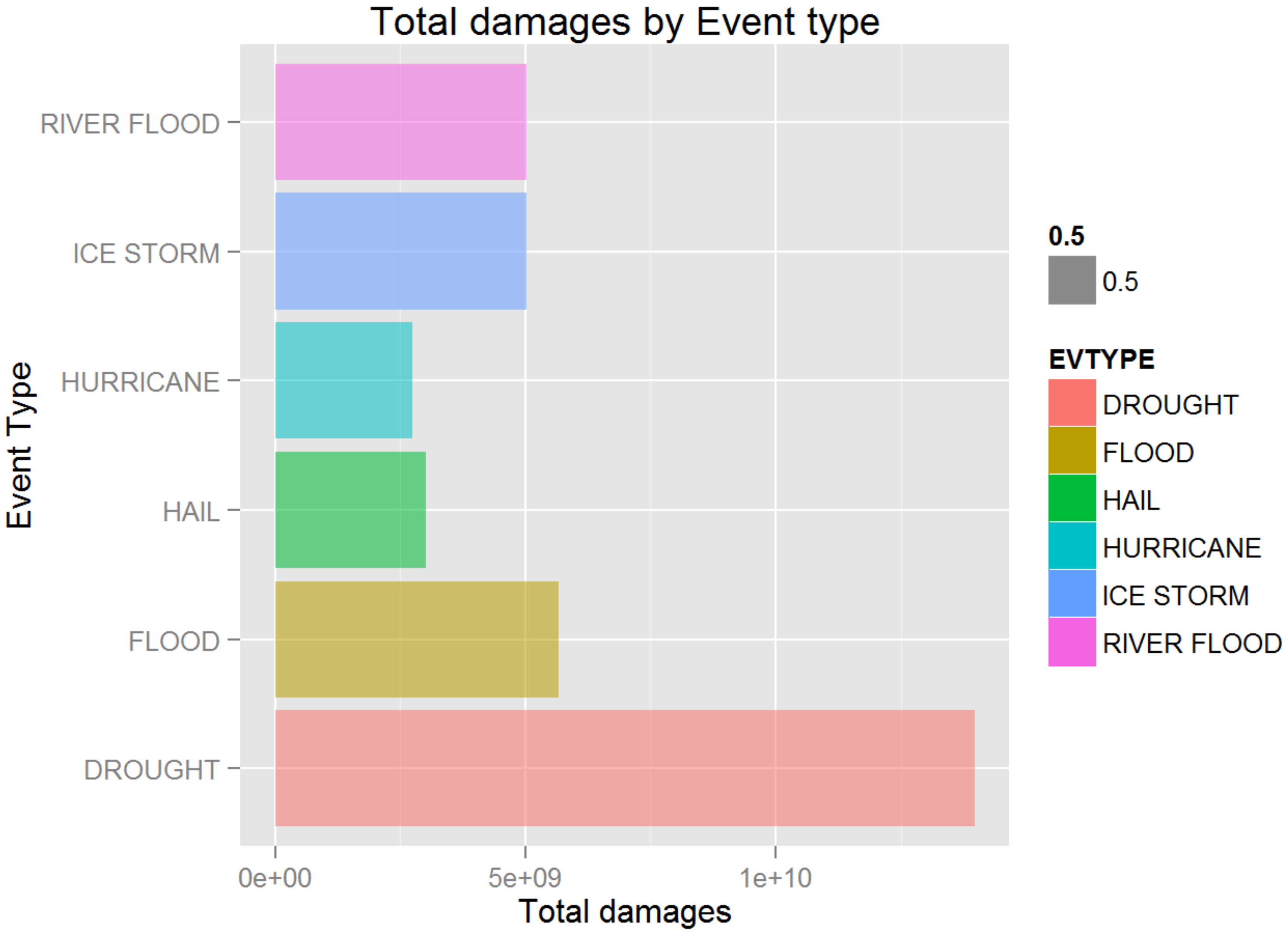
Now we make a new dataset of crop damage arranged according to events type and look at the first 6 major events in terms of economic loss.

```
Damage.crop = ddply(data, .(EVTYPE), summarize, damage.crop = sum(damage.crop, na.rm = TRUE
))
# Sort the Damage.crop dataset
Damage.crop = Damage.crop[order(Damage.crop$damage.crop, decreasing = T), ]
# Show the first 6 most damaging types
head(Damage.crop)
```

```
##          EVTYPE damage.crop
## 95      DROUGHT   1.397e+10
## 170       FLOOD   5.662e+09
## 590 RIVER FLOOD   5.029e+09
## 427   ICE STORM   5.022e+09
## 244        HAIL   3.026e+09
## 402   HURRICANE   2.742e+09
```

Let’s also look at a chart to have a quick look at these figures.

```
ggplot(Damage.crop[1:6, ], aes(EVTYPE, damage.crop, fill = EVTYPE, alpha=0.5)) + geom_bar(stat = "identity") +
  xlab("Event Type") + ylab("Total damages") + ggtitle("Total damages by Event type") + coord_flip()
```



We see that drought is the worst factor for agriculture causing more than 13 billion dollars. This is followed by flood causing more than 5 billion dollars.

Now if we want to look at the econmoic losses at aggregate, we need to add the losses from property and crop and then look at which type is most devastating, flood or drought.

Let’s compute total damage first and combine it to data.Then we just need to segregate losses according to event types.

```
total.damage = damage.property + damage.crop
data=as.data.frame(cbind(data,total.damage))
Damage.total = ddply(data, .(EVTYPE), summarize, damage.total = sum(total.damage, na.rm = T
RUE))
# Sort the Damage.crop dataset
Damage.total = Damage.total[order(Damage.total$damage.total, decreasing = T), ]
```

Let’s have a look at first 6 most damaging types

```
head(Damage.total)
```

```
##                EVTYPE damage.total
## 170             FLOOD    1.503e+11
## 411 HURRICANE/TYPHOON    7.191e+10
## 834           TORNADO    5.736e+10
## 670       STORM SURGE    4.332e+10
## 244              HAIL    1.876e+10
## 153       FLASH FLOOD    1.824e+10
```

We see that it is flood with a whooping loss of more than 150 billions followed by hurricane/typhoon with an estimate of more than 71 billions.In terms of total losses, drought–main economic loss event is not even among the loss inducing six events.

Result

It is evident from the exploratory analysis presented here that flood is the most exacerbating factor for economic loss while tornado is for population health. If agriculture is the main concern, then drought may be the most concerning factor for economy. But for the economy in general, flood becomes the main loss factor. Concerning human health, more priorities naturally will go towards addressing tornado as it claims most lives and causes injuries.