Analysis of the impact of severe weather on property and crops

Reproducible research Assignment 2

Synopsis

This report analyses the NOAA Storm Database to answer the following questions about severe weather data events:

1. Across the United States, which types of events are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

Data is downloaded from here (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2)

Information about the data is available at:

- National Weather Service Storm Data Documentation (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf) and

- National Climatic Data Center Storm Events FAQ (https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2FNCDC%20Storm%20Events-FAQ%20Page.pdf)

An exploratory analysis has been done on the data to have a look at the dimensions, variable names and values of the factor variables; data related to the variables in which we are interested (i.e. event type, damage to health which is represented by injuries and fatalities and property and crop damage which represent economic loss) has been cleaned and pre-processed in order to be suitable to reply to the given questions.

Data has subsequently been subsetted to the variables of interest for the questions and aggregated in order to count the incidents and damage caused by each type of event and the ten highest values for both questions are presented in the results section together with a plot.

Data Processing

In this section the processing done on the data is described starting from the downloading.

Downloading and loading the data

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

Exploratory analysis on the data

Checking how many data we have, i.e. how many observations and how many event types are recorded in the database.

```
## Total number of observations in the database
dim(storm)
```

```
## [1] 902297     37
```

```
## Total number of different events types recorded in the database
length(unique(storm$EVTYPE))
```

```
## [1] 985
```

The number of different events recorded is quite high.

From a quick look at the EVTYPE variable we see that names are mispelled or the same event is recorded with a different name (for example THUNDERSTORM WIND is recorded als also as THUNDERSTORM WINS and THUNDERSTORM WINDS).

```
## display the value of EVTYPE in the first 20 observations
head(unique(storm$EVTYPE),n=20)
```

```
##  [1] TORNADO                   TSTM WIND
##  [3] HAIL                      FREEZING RAIN
##  [5] SNOW                      ICE STORM/FLASH FLOOD
##  [7] SNOW/ICE                  WINTER STORM
##  [9] HURRICANE OPAL/HIGH WINDS THUNDERSTORM WINDS
## [11] RECORD COLD               HURRICANE ERIN
## [13] HURRICANE OPAL            HEAVY RAIN
## [15] LIGHTNING                 THUNDERSTORM WIND
## [17] DENSE FOG                 RIP CURRENT
## [19] THUNDERSTORM WINS         FLASH FLOOD
## 985 Levels:    HIGH SURF ADVISORY  COASTAL FLOOD ... WND
```

The following functions will help in transforming the names in order to group the same events labelled differently. Names are shortened and simplified.

```
trim <- function( x ) {
  gsub("(^ +)|( +$)", "", x)
}

clean_EVTYPE <- function(data) {
  data$EVTYPE <- trim(as.character(data$EVTYPE))
  data$EVTYPE  <- tolower(data$EVTYPE)
  data$EVTYPE[grep("thun|tstm|ligh", data$EVTYPE)] <- "thunderstorm"
  data$EVTYPE[grep("heat", data$EVTYPE)] <- "heat"
  data$EVTYPE[grep("avalance", data$EVTYPE)] <- "avalanche"
  data$EVTYPE[grep("fire", data$EVTYPE)] <- "fire"
  data$EVTYPE[grep("wint", data$EVTYPE)] <- "winter weather"
  data$EVTYPE[grep("snow", data$EVTYPE)] <- "snow"
  data$EVTYPE[grep("flood|stream fld", data$EVTYPE)] <- "flood"
  data$EVTYPE[grep("rain", data$EVTYPE)] <- "rain"
  data$EVTYPE[grep("cold", data$EVTYPE)] <- "cold"
  data$EVTYPE[grep("cur", data$EVTYPE)] <- "rip current"
  data$EVTYPE <- factor(data$EVTYPE)
  data
}
```

Similarly, a quick look at the variables holding data about the amount of damage to property and crops reveals that they are stored with heterogeneous terms. For example “Million” is indicated as “m”, “M” or as 6 digits.

```
## display the values of the variables containing the exponent for property and crop damag
e amounts
unique(storm$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
```

```
unique(storm$CROPDMGEXP)
```

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

The following functions will help in transforming the data regarding the amount of damage to property and crops in a form that can be easily used to do math.

```
cvt_PROPDMGEXP <- function(data) {
  data$PROPDMGEXP <- as.character(data$PROPDMGEXP)
  data$PROPDMGEXP <- trim(data$PROPDMGEXP)
  data$PROPDMGEXP <- tolower(data$PROPDMGEXP)
  data$PROPDMGEXP[data$PROPDMGEXP == "k" | data$PROPDMGEXP == "3"] <- 1000
  data$PROPDMGEXP[data$PROPDMGEXP == "m" | data$PROPDMGEXP == "6"] <- 1000000
  data$PROPDMGEXP[data$PROPDMGEXP == "b"] <- 1000000000
  data$PROPDMGEXP[data$PROPDMGEXP == "h" | data$PROPDMGEXP == "2"] <- 100
  data$PROPDMGEXP[data$PROPDMGEXP == "+" | data$PROPDMGEXP == "-" | data$PROPDMGEXP == "" | data$PROPDMGEXP == "0"] <- 1
  data$PROPDMGEXP[data$PROPDMGEXP == "4"] <- 10000
  data$PROPDMGEXP[data$PROPDMGEXP == "5"] <- 100000
  data$PROPDMGEXP[data$PROPDMGEXP == "7"] <- 10000000
  data$PROPDMGEXP <- as.numeric(data$PROPDMGEXP)
  data
}

cvt_CROPDMGEXP <- function(data) {
  data$CROPDMGEXP <- as.character(data$CROPDMGEXP)
  data$CROPDMGEXP <- trim(data$CROPDMGEXP)
  data$CROPDMGEXP <- tolower(data$CROPDMGEXP)
  data$CROPDMGEXP[data$CROPDMGEXP == "2"] <- 100
  data$CROPDMGEXP[data$CROPDMGEXP == "k"] <- 1000
  data$CROPDMGEXP[data$CROPDMGEXP == "m"] <- 1000000
  data$CROPDMGEXP[data$CROPDMGEXP == "b"] <- 1000000000
  data$CROPDMGEXP[data$CROPDMGEXP == "" | data$CROPDMGEXP == "0" | data$CROPDMGEXP == "?"] <- 1
  data$CROPDMGEXP <- as.numeric(data$CROPDMGEXP)
  data
}
```

Subsetting and aggregating the data to extract the information of interest

To reply to the first question on harmfulness of the events, the data related to the type of event, and incidents are preprocessed in this part. Incidents are represented in the database as injuries and fatalities. So the harmfulness of the event will be expressed in terms of injuries plus fatalities caused by each event.

```
s_injuries<-storm[storm$INJURIES > 0 | storm$FATALITIES > 0, c("EVTYPE", "INJURIES", "FATALITIES")]
s_injuries<-clean_EVTYPE(s_injuries)
s_injuries$INCIDENTS <- s_injuries$FATALITIES + s_injuries$INJURIES
```

Data are aggregated by summing on the type of event, then reordered in descending order by the amount of incidents.

```
s_incidents <-aggregate(list(FATALITIES=s_injuries$FATALITIES, INJURIES=s_injuries$INJURIE
S, INCIDENTS=s_injuries$INCIDENTS),
    by=list(s_injuries$EVTYPE), FUN=sum)
names(s_incidents)[1] <- "EVENT_TYPE"
s_incidents <- s_incidents[order(-s_incidents$INCIDENTS), ]
```

The ten most harmful events are extracted from the list of incidents and re-enumerated in order to have them ordered in terms of damage and not in alphabetical order per event type.

```
top10_incidents <- head(s_incidents,n=10)
top10_incidents$EVENT_TYPE <- factor(top10_incidents$EVENT_TYPE,
  levels=top10_incidents$EVENT_TYPE[order(-top10_incidents$INCIDENTS)])
row.names(top10_incidents) <- 1:nrow(top10_incidents)
```

To reply to the second question on the event which has the worst economic consequences, the data related to the type of event, production damage and crop damage are preprocessed in this part. Data related to damage is contained in four variables: * PROPDMG (property damage) * PROPDMGEXP (property damage exponent) * CROPDMG (crop damage) * CROPDMGEXP (crop damage exponent)

Exponents are normalised and data transformed in numeric values for math operations.

```
s_damage <- storm[storm$PROPDMG != 0 | storm$CROPDMG != 0, c("EVTYPE", "PROPDMG", "PROPDMG
EXP", "CROPDMG", "CROPDMGEXP")]
s_damage <- clean_EVTYPE(s_damage)
s_damage <- cvt_PROPDMGEXP(s_damage)
s_damage <- cvt_CROPDMGEXP(s_damage)
s_damage$PROPDMG<-s_damage$PROPDMG*s_damage$PROPDMGEXP
s_damage$CROPDMG<-s_damage$CROPDMG*s_damage$CROPDMGEXP
s_damage$TOT_DMG <- s_damage$PROPDMG + s_damage$CROPDMG
```

Data are aggregated by summing on the event type.

```
s_tot_damage <- aggregate(list(PROPERTY_DMG=s_damage$PROPDMG, CROP_DMG=s_damage$CROPDMG, T
OT_DMG=s_damage$TOT_DMG),
                          list(s_damage$EVTYPE), FUN=sum)
names(s_tot_damage)[1] <- "EVENT_TYPE"
s_tot_damage <- s_tot_damage[order(-s_tot_damage$TOT_DMG), ]
```

The ten events with highest economic consequences are extracted and re-enumerated in order to have them ordered in terms of damage and not in alphabetical order per event type.

```
top10_damage <- head(s_tot_damage,n=10)
top10_damage$EVENT_TYPE <- factor(top10_damage$EVENT_TYPE, levels=top10_damage$EVENT_TYPE[
order(-top10_damage$TOT_DMG)])
row.names(top10_damage) <- 1:nrow(top10_damage)
```

Results

The most harmful events in terms of population health are:

```
top10_incidents
```

```
##           EVENT_TYPE FATALITIES INJURIES INCIDENTS
## 1            tornado       5633    91346     96979
## 2       thunderstorm       1574    14778     16352
## 3               heat       3138     9224     12362
## 4              flood       1553     8683     10236
## 5     winter weather        279     1968      2247
## 6          ice storm         89     1975      2064
## 7               fire         90     1608      1698
## 8          high wind        248     1137      1385
## 9               hail         15     1361      1376
## 10 hurricane/typhoon         64     1275      1339
```

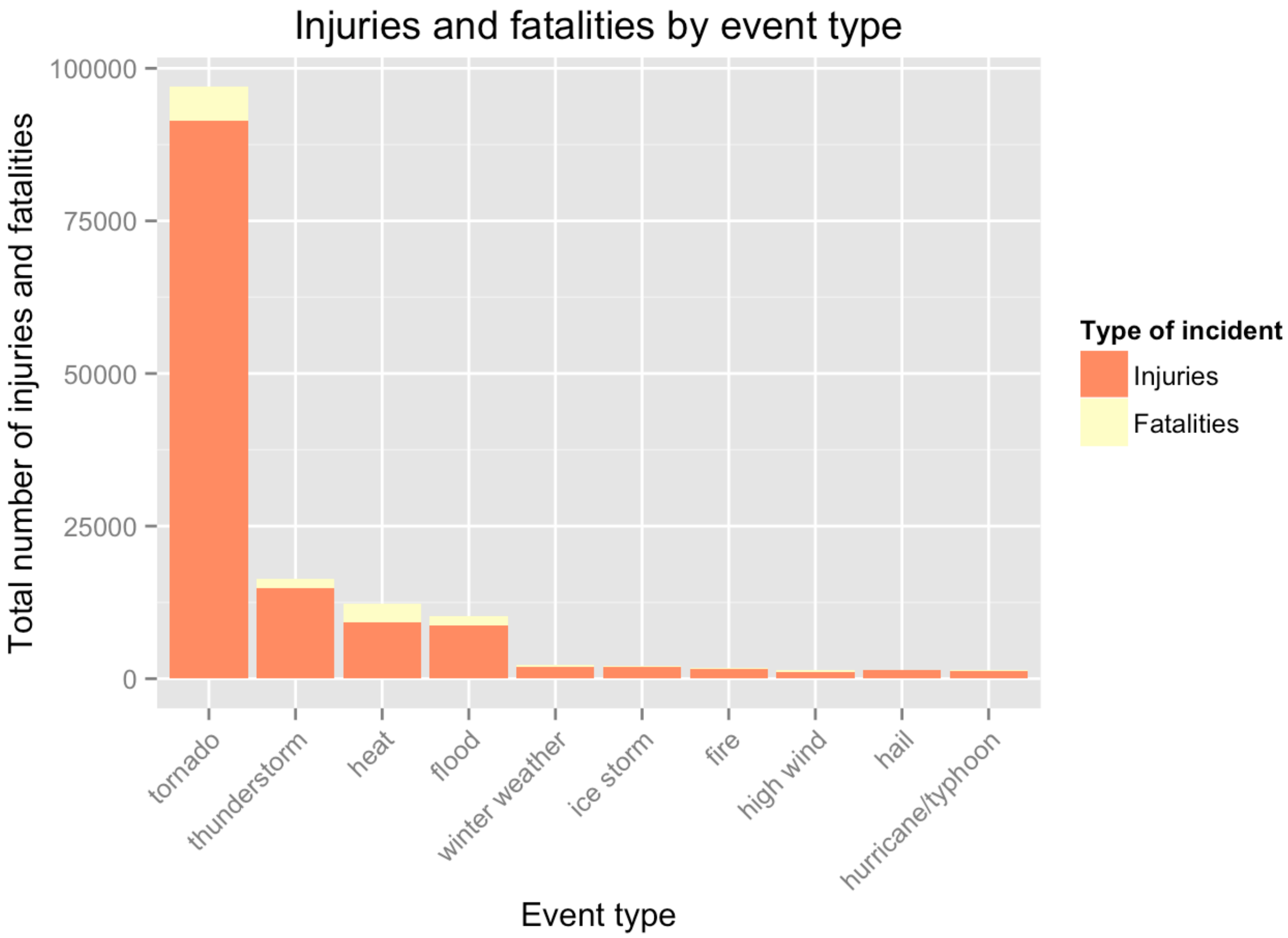
The ten events with the greatest economic consequences are:

```
top10_damage
```

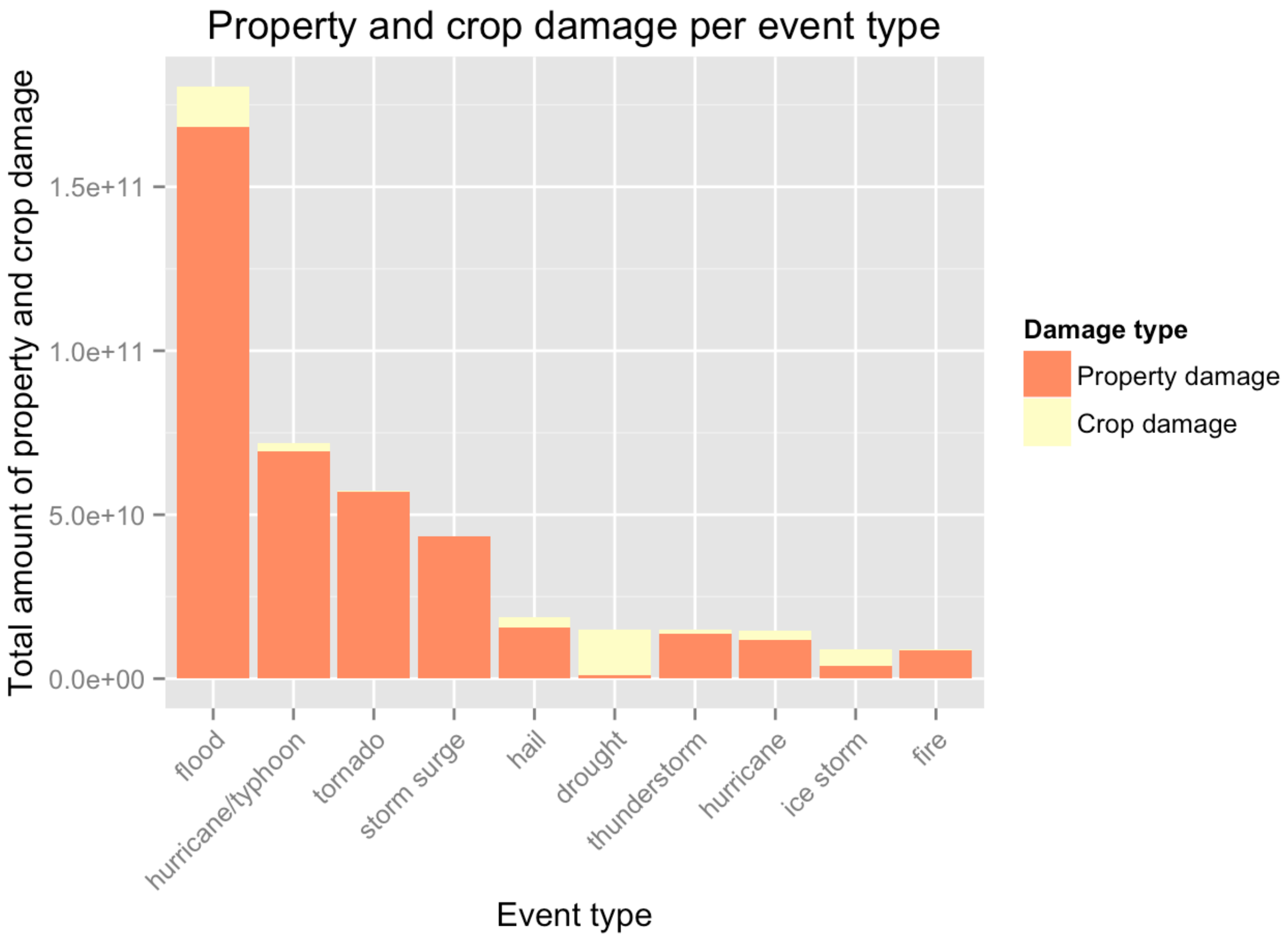
```
##           EVENT_TYPE PROPERTY_DMG  CROP_DMG   TOT_DMG
## 1              flood    1.683e+11 1.239e+10 1.807e+11
## 2  hurricane/typhoon    6.931e+10 2.608e+09 7.191e+10
## 3            tornado    5.695e+10 4.150e+08 5.736e+10
## 4        storm surge    4.332e+10 5.000e+03 4.332e+10
## 5               hail    1.574e+10 3.026e+09 1.876e+10
## 6            drought    1.046e+09 1.397e+10 1.502e+10
## 7       thunderstorm    1.373e+10 1.286e+09 1.502e+10
## 8          hurricane    1.187e+10 2.742e+09 1.461e+10
## 9          ice storm    3.945e+09 5.022e+09 8.967e+09
## 10              fire    8.497e+09 4.033e+08 8.900e+09
```

The results are graphycally displayed by the plots below:

```
library(ggplot2)
library(reshape2)

incidents_plot <- melt(top10_incidents[c("EVENT_TYPE", "INJURIES", "FATALITIES")],id.vars
= "EVENT_TYPE")

ggplot(incidents_plot, aes(x = EVENT_TYPE, y = value, fill=variable)) +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(name="Type of incident",  breaks=c("INJURIES", "FATALITIES"),
                         labels=c("Injuries", "Fatalities"), palette="Spectral") +
  labs(title=expression("Injuries and fatalities by event type"),
       y=expression("Total number of injuries and fatalities"),
       x="Event type")
```



```
incidents_plot$value <-log10(incidents_plot$value)

damage_plot <- melt(top10_damage[c("EVENT_TYPE", "PROPERTY_DMG", "CROP_DMG")],id.vars = "E
VENT_TYPE")

ggplot(damage_plot, aes(x = EVENT_TYPE, y = value, fill=variable)) +
  geom_bar(stat="identity") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(name="Damage type", breaks=c("PROPERTY_DMG", "CROP_DMG"),
                       labels=c("Property damage", "Crop damage"), palette="Spectral") +
  labs(title=expression("Property and crop damage per event type"),
       y=expression("Total amount of property and crop damage"),
       x="Event type")
```



Conclusions

From the table and plots presented in the results section it can be concluded that:

- tornadoes and thunderstorms are the events most harmful with respect to population health
- floods and hurricane/typhoons are the events have the greatest economic consequences on property and crops.