Coursera Reproducible Research Project 2

Antoine Mertz

2017-05-11

Synopsis

This analysis is the final project of the Coursera Reproducible Research course, part of the Data Science Specialization. The purpose of the work is to explore the NOAA Storm Database and explore the effects of severe weather events on population health and economy.

The analysis goal is to find which types of sever weather events are most harmful on the population health (injuries and fatalities) and have the greatest economic consequences (property and crop damages).

Data Processing

The data can be downloaded in the project assignment page in coursera web site. There is also some documentation:

- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQ

During the analysis, these following packages will be necessary

```
library(dplyr, warn.conflicts = FALSE)
library(ggplot2, warn.conflicts = FALSE)
library(gridExtra, warn.conflicts = FALSE)
```

First we need to read the data, assuming that the file is downloaded in your current working directory.

```
# read data
storm.df <- read.csv(bzfile("repdata%2Fdata%2FStormData.csv.bz2"))
```

Then we subset the dataset, because, we don't need all the variables available for the analysis (we just keep `EVTYPE`, `FATALITIES`, `INJURIES`, `PROPDMG`, `PROPDMGEXP`, `CROPDMG` and `CROPDMGEXP`).

```
# subset data
sub.storm.df <- storm.df %>%
  select(EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP)
```

Then we process the data to calculate the property and crop damages value.

```
# calculate Property Damages
sub.storm.df <- sub.storm.df %>%
  mutate(PROPDMGEXP = toupper(PROPDMGEXP)) %>%
  mutate(PROPERTYDAMAGE = ifelse(PROPDMGEXP == "H", PROPDMG * 1e2,
                                 ifelse(PROPDMGEXP == "K", PROPDMG * 1e3,
                                        ifelse(PROPDMGEXP == "M", PROPDMG * 1e6,
                                               ifelse(PROPDMGEXP == "B", PROPDMG * 1e9, 0)))))

# calculate Crop Damages
sub.storm.df <- sub.storm.df %>%
  mutate(CROPDMGEXP = toupper(CROPDMGEXP)) %>%
```

```
  mutate(CROPDAMAGE = ifelse(CROPDMGEXP == "H", CROPDMG * 1e2,
                             ifelse(CROPDMGEXP == "K", CROPDMG * 1e3,
                                    ifelse(CROPDMGEXP == "M", CROPDMG * 1e6,
                                           ifelse(CROPDMGEXP == "B", CROPDMG * 1e9, 0)))))
```

Results

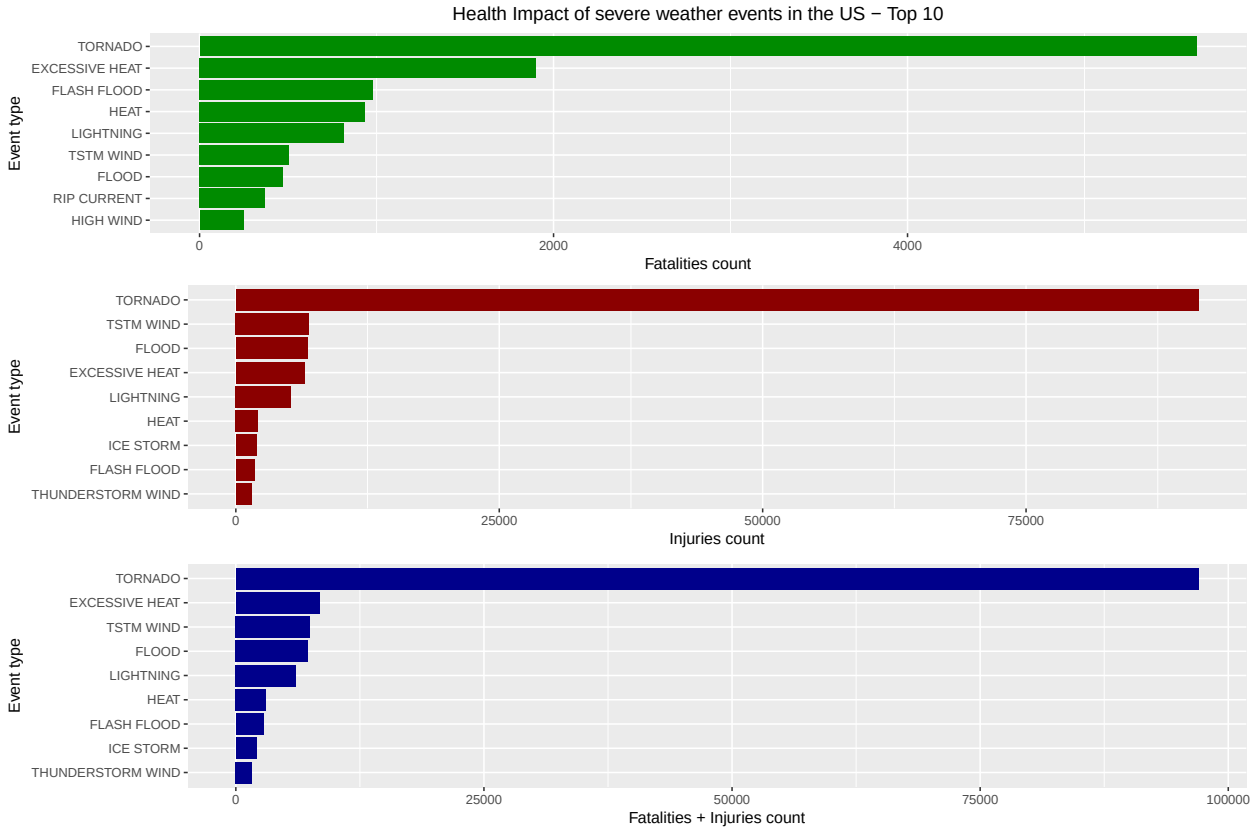
Public health impacts

The first question we have to answer for the project is "**Across the United States, which types of events (as indicated in the `EVTYPE` variable) are most harmful with respect to population health?**"

We can represent the number of fatalities and injuries by events to display the top 10 most harmful (in one figure within 3 plots using a panel plots).

```
# count fatalities by event type
fatalities <- sub.storm.df %>%
  group_by(EVTYPE) %>%
  summarize(COUNT = sum(FATALITIES, na.rm = TRUE)) %>%
  arrange(desc(COUNT)) %>%
  as.data.frame()

fatalities$type <- "fatalities"

# count fatalities by event type
injuries <- sub.storm.df %>%
  group_by(EVTYPE) %>%
  summarize(COUNT = sum(INJURIES, na.rm = TRUE)) %>%
  arrange(desc(COUNT)) %>%
  as.data.frame()

injuries$type <- "injuries"

# merge fatalities and injuries by event type
pop.health <- rbind(fatalities, injuries)

# first plot the top10 event types with the more fatalities
g1 <- pop.health %>%
  filter(type == "fatalities") %>%
  arrange(desc(COUNT)) %>%
  filter(row_number() < 10) %>%
  ggplot(aes(x=reorder(EVTYPE, COUNT), y=COUNT)) +
  geom_bar(stat = "identity", fill = "green4") +
  xlab("Event type") +
  ylab("Fatalities count") +
  ggtitle("Health Impact of severe weather events in the US - Top 10") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_flip()

# second plot the top10 event types with the more injuries
g2 <- pop.health %>%
  filter(type == "injuries") %>%
```

```
  arrange(desc(COUNT)) %>%
  filter(row_number() < 10) %>%
  ggplot(aes(x=reorder(EVTYPE, COUNT), y=COUNT)) +
  geom_bar(stat = "identity", fill = "red4") +
  xlab("Event type") +
  ylab("Injuries count") +
  coord_flip()

# finally plot the top10 event types the most harmful (injuries and fatalities mixed)
g3 <- pop.health %>%
  group_by(EVTYPE) %>%
  summarise(TOTAL_COUNT = sum(COUNT, na.rm = TRUE)) %>%
  arrange(desc(TOTAL_COUNT)) %>%
  filter(row_number() < 10) %>%
  ggplot(aes(x=reorder(EVTYPE, TOTAL_COUNT), y=TOTAL_COUNT)) +
  geom_bar(stat = "identity", fill = "blue4") +
  xlab("Event type") +
  ylab("Fatalities + Injuries count") +
  coord_flip()

# make a panel plots
grid.arrange(g1, g2, g3, nrow = 3)
```



According to the plot, tornados are by far the most dangerous severe weather event for population health. Heat, floods and wind also seem to cause a high number of deaths and injuries.

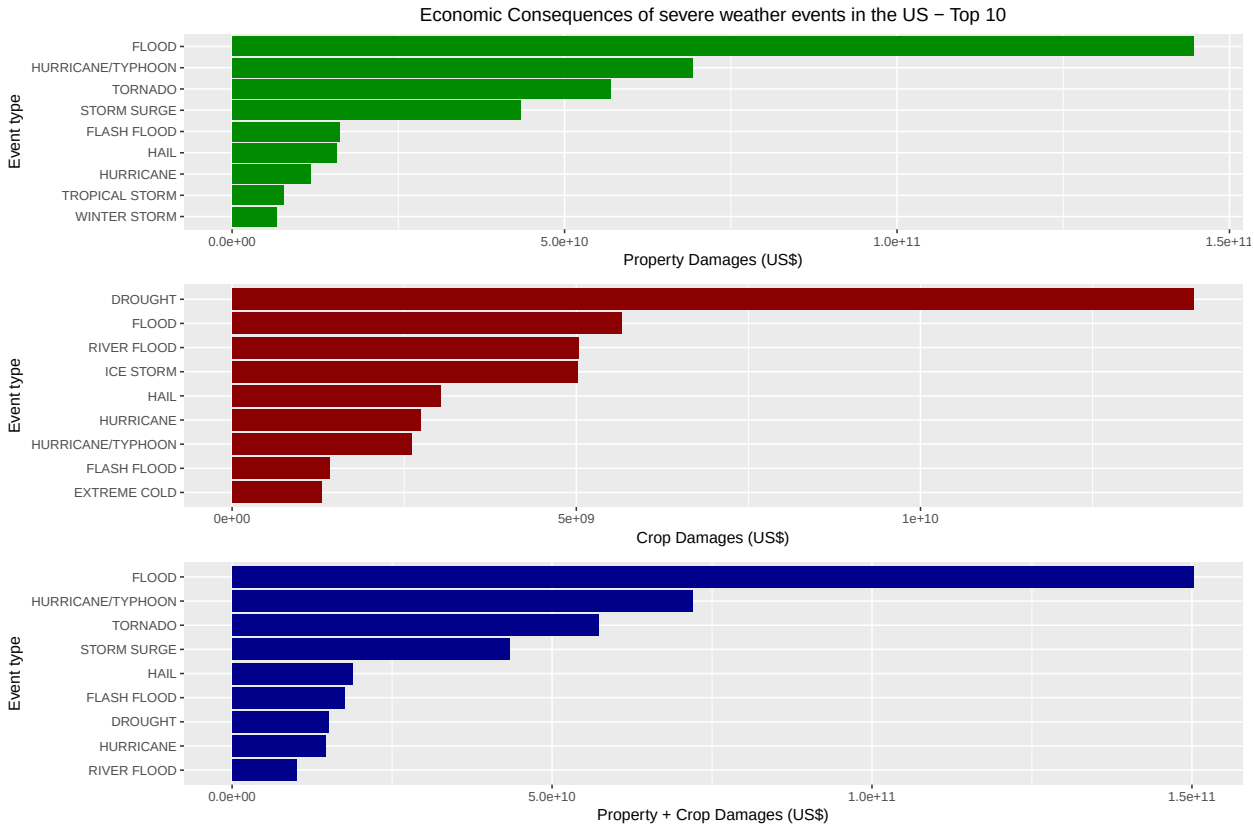
Economic impacts

The second question we have to answer for the project is "**Across the United States, which types of events have the greatest economic consequences?**"

Like for public health impacts, we can represent the amount of property and corp damages by events to display the top 10 most impactful (in one figure within 3 plots using a panel plots).

```
# calculate property damages by event type
prop.damages <- sub.storm.df %>%
  group_by(EVTYPE) %>%
  summarize(SUM_DAMAGES = sum(PROPERTYDAMAGE, na.rm = TRUE)) %>%
  as.data.frame()

prop.damages$type <- "prop"

# calculate crop damages by event type
crop.damages <- sub.storm.df %>%
  group_by(EVTYPE) %>%
  summarize(SUM_DAMAGES = sum(CROPDAMAGE, na.rm = TRUE)) %>%
  as.data.frame()

crop.damages$type <- "crop"

# merge crop and property damages
damages <- rbind(prop.damages, crop.damages)

# first plot the top10 event types with the more impact on property economy
g1 <- damages %>%
  filter(type == "prop") %>%
  arrange(desc(SUM_DAMAGES)) %>%
  filter(row_number() < 10) %>%
  ggplot(aes(x=reorder(EVTYPE, SUM_DAMAGES), y=SUM_DAMAGES)) +
  geom_bar(stat = "identity", fill = "green4") +
  xlab("Event type") +
  ylab("Property Damages (US$)") +
  ggtitle("Economic Consequences of severe weather events in the US - Top 10") +
  theme(plot.title = element_text(hjust = 0.5)) +
  coord_flip()

# second plot the top10 event types with the more impact on crop economy
g2 <- damages %>%
  filter(type == "crop") %>%
  arrange(desc(SUM_DAMAGES)) %>%
  filter(row_number() < 10) %>%
  ggplot(aes(x=reorder(EVTYPE, SUM_DAMAGES), y=SUM_DAMAGES)) +
  geom_bar(stat = "identity", fill = "red4") +
  xlab("Event type") +
  ylab("Crop Damages (US$)") +
  coord_flip()

# finally plot the top10 event types with the more economic consequences
g3 <- damages %>%
  group_by(EVTYPE) %>%
  summarise(TOTAL_DAMAGES = sum(SUM_DAMAGES, na.rm = TRUE)) %>%
```

```
  arrange(desc(TOTAL_DAMAGES)) %>%
  filter(row_number() < 10) %>%
  ggplot(aes(x=reorder(EVTYPE, TOTAL_DAMAGES), y=TOTAL_DAMAGES)) +
  geom_bar(stat = "identity", fill = "blue4") +
  xlab("Event type") +
  ylab("Property + Crop Damages (US$)") +
  coord_flip()

# make a panel plots
grid.arrange(g1, g2, g3, nrow = 3)
```



Flood is, by far, the severe weather event with the more consequences on economy. Typhoons and tornado cost also a lot. In term of crop damages, floods, river floods and ice storms have a lot of consequences.