Reproducible Research Course Project 2

Analysis of the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database

This project explores the NOAA storm database, which tracks major storms and weather events, to address the most severe types of weather events in the USA, which caused greatest damage to human population in terms of fatalities/injuries and economic loss during the years 1950 - 2011.

There are two goals of this analysis:

- identify the weather events that are most harmful with respect to population health
- identify the weather events that have the greatest economic consequences.

Based on our analysis, we conclude that TORNADOS and FLOODS are most harmful weather events in the USA in terms of the risk to human health and economic impact.

Data Processing

```
# downloading data
Url_data <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
File_data <- "StormData.csv.bz2"
if (!file.exists(File_data)) {
        download.file(Url_data, File_data, mode = "wb")
}
# reading data
Raw_data <- read.csv(file = File_data, header=TRUE, sep=",")
```

The data source is in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. It is possible to download the source file from the course web site: Storm Data

Additional documentation on the database was provided here:

- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQ
- Mentor's comments in the Discussion Forum on the Course web-site

```
# subsetting by date
Main_data <- Raw_data
Main_data$BGN_DATE <- strptime(Raw_data$BGN_DATE, "%m/%d/%Y %H:%M:%S")
Main_data <- subset(Main_data, BGN_DATE > "1995-12-31")
```

1. According to NOAA, the data recording start from Jan. 1950. At that time, they recorded only one event type - tornado. They added more events gradually, and only from Jan 1996 they started recording all events type. Since our objective is comparing the effects of different weather events, we need only to include events that started not earlier than Jan 1996.

2. Based on the above mentioned documentation and preliminary exploration of raw data with 'str', 'names', 'table', 'dim', 'head', 'range' and other similar functions we can conclude that there are 7 variables we are interested in regarding the two questions.

Namely: EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP.

```
Main_data <- subset(Main_data, select = c(EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CI
```

Therefore, we can limit our data to these variables.

Contents of data now are as follows: EVTYPE – type of event
FATALITIES – number of fatalities
INJURIES – number of injuries
PROPDMG – the size of property damage
PROPDMGEXP - the exponent values for 'PROPDMG' (property damage)
CROPDMG - the size of crop damage
CROPDMGEXP - the exponent values for 'CROPDMG' (crop damage)

```
#cleaning event types names
Main_data$EVTYPE <- toupper(Main_data$EVTYPE)
# eliminating zero data
Main_data <- Main_data[Main_data$FATALITIES !=0 |
                       Main_data$INJURIES !=0 |
                       Main_data$PROPDMG !=0 |
                       Main_data$CROPDMG !=0, ]
```

3. There are almost 1000 unique event types in EVTYPE column. Therefore, it is better to limit database to a reasonable number. We can make it by capitalizing all letters in EVTYPE column as well as subsetting only non-zero data regarding our target numbers.

Now we have 186 unique event types and it seems like something to work with.

Population health data processing

```
Health_data <- aggregate(cbind(FATALITIES, INJURIES) ~ EVTYPE, data = Main_data, FUN=sum)
Health_data$PEOPLE_LOSS <- Health_data$FATALITIES + Health_data$INJURIES
Health_data <- Health_data[order(Health_data$PEOPLE_LOSS, decreasing = TRUE), ]
Top10_events_people <- Health_data[1:10,]
knitr::kable(Top10_events_people, format = "markdown")
```

We aggregate fatalities and injuries numbers in order to identify TOP-10 events contributing the total people loss:

	EVTYPE	FATALITIES	INJURIES	PEOPLE_LOSS
149	TORNADO	1511	20667	22178
39	EXCESSIVE HEAT	1797	6391	8188
48	FLOOD	414	6758	7172
107	LIGHTNING	651	4141	4792
153	TSTM WIND	241	3629	3870
46	FLASH FLOOD	887	1674	2561
146	THUNDERSTORM WIND	130	1400	1530
182	WINTER STORM	191	1292	1483
69	HEAT	237	1222	1459
88	HURRICANE/TYPHOON	64	1275	1339

Economic consequences data processing

The number/letter in the exponent value columns (PROPDMGEXP and CROPDMGEXP) represents the power of ten (10^The number). It means that the total size of damage is the product of PROPDMG and CROPDMG and figure 10 in the power corresponding to exponent value.

Exponent values are:

- numbers from one to ten
- letters (B or b = Billion, M or m = Million, K or k = Thousand, H or h = Hundred)
- and symbols "-", "+" and "?" which refers to less than, greater than and low certainty. We have the option to ignore these three symbols altogether.

```
Main_data$PROPDMGEXP <- gsub("[Hh]", "2", Main_data$PROPDMGEXP)
Main_data$PROPDMGEXP <- gsub("[Kk]", "3", Main_data$PROPDMGEXP)
Main_data$PROPDMGEXP <- gsub("[Mm]", "6", Main_data$PROPDMGEXP)
Main_data$PROPDMGEXP <- gsub("[Bb]", "9", Main_data$PROPDMGEXP)
Main_data$PROPDMGEXP <- gsub("\\+", "1", Main_data$PROPDMGEXP)
Main_data$PROPDMGEXP <- gsub("\\?|\\-|\\ ", "0",  Main_data$PROPDMGEXP)
Main_data$PROPDMGEXP <- as.numeric(Main_data$PROPDMGEXP)
Main_data$CROPDMGEXP <- gsub("[Hh]", "2", Main_data$CROPDMGEXP)
Main_data$CROPDMGEXP <- gsub("[Kk]", "3", Main_data$CROPDMGEXP)
Main_data$CROPDMGEXP <- gsub("[Mm]", "6", Main_data$CROPDMGEXP)
Main_data$CROPDMGEXP <- gsub("[Bb]", "9", Main_data$CROPDMGEXP)
```

```
Main_data$CROPDMGEXP <- gsub("\\+", "1", Main_data$CROPDMGEXP)
Main_data$CROPDMGEXP <- gsub("\\-|\\?|\\ ", "0", Main_data$CROPDMGEXP)
Main_data$CROPDMGEXP <- as.numeric(Main_data$CROPDMGEXP)
Main_data$PROPDMGEXP[is.na(Main_data$PROPDMGEXP)] <- 0
Main_data$CROPDMGEXP[is.na(Main_data$CROPDMGEXP)] <- 0
```

We transform letters and symbols to numbers:

```
#creating total damage values
library(dplyr)
Main_data <- mutate(Main_data,
                    PROPDMGTOTAL = PROPDMG * (10 ^ PROPDMGEXP),
                    CROPDMGTOTAL = CROPDMG * (10 ^ CROPDMGEXP))
```

At last, we create new values of total property damage and total crop damage for analysis (we need 'dplr' package for that).

```
Economic_data <- aggregate(cbind(PROPDMGTOTAL, CROPDMGTOTAL) ~ EVTYPE, data = Main_data, FUN=sum)
Economic_data$ECONOMIC_LOSS <- Economic_data$PROPDMGTOTAL + Economic_data$CROPDMGTOTAL
Economic_data <- Economic_data[order(Economic_data$ECONOMIC_LOSS, decreasing = TRUE), ]
Top10_events_economy <- Economic_data[1:10,]
knitr::kable(Top10_events_economy, format = "markdown")
```

Now we aggregate property and crop damage numbers in order to identify TOP-10 events contributing the total economic loss:

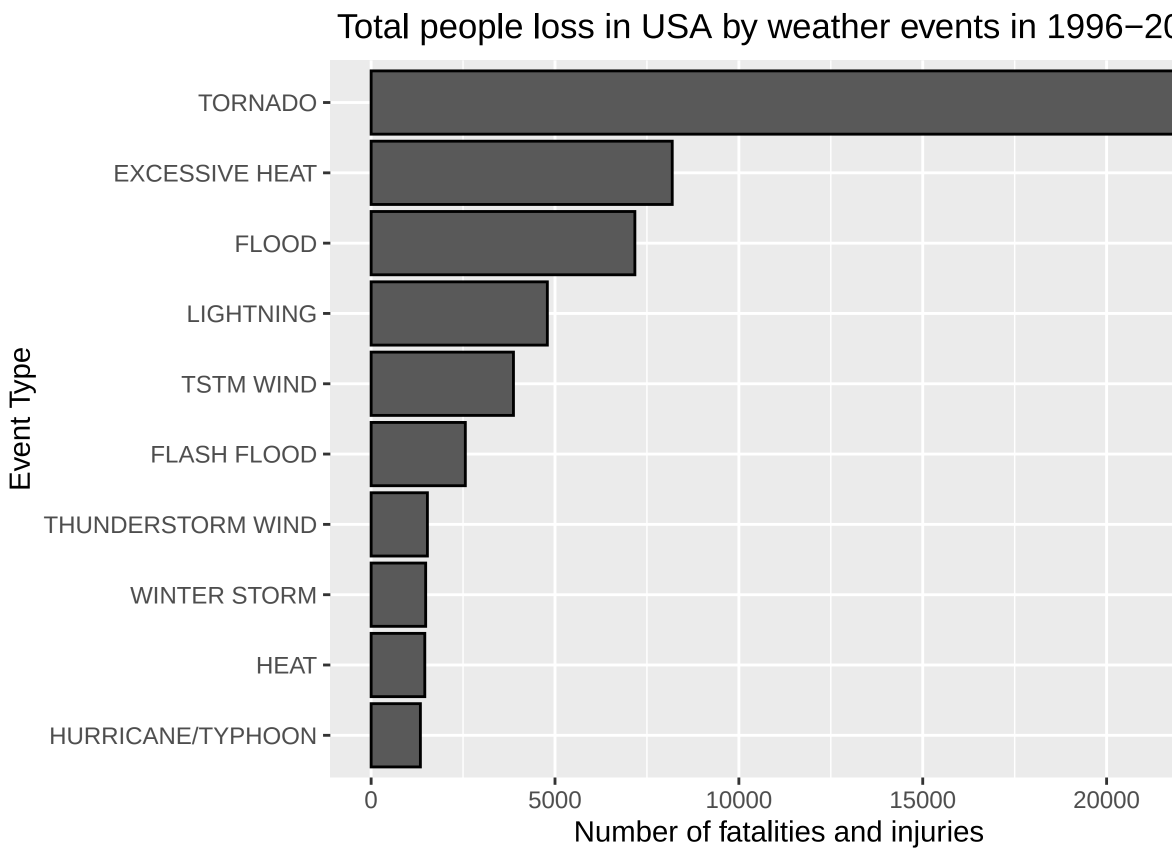
	EVTYPE	PROPDMGTOTAL	CROPDMGTOTAL	ECONOMIC_LOSS
48	FLOOD	143944833550	4974778400	148919611950
88	HURRICANE/TYPHOON	69305840000	2607872800	71913712800
141	STORM SURGE	43193536000	5000	43193541000
149	TORNADO	24616945710	283425010	24900370720
66	HAIL	14595143420	2476029450	17071172870
46	FLASH FLOOD	15222203910	1334901700	16557105610
86	HURRICANE	11812819010	2741410000	14554229010
32	DROUGHT	1046101000	13367566000	14413667000
152	TROPICAL STORM	7642475550	677711000	8320186550
83	HIGH WIND	5247860360	633561300	5881421660

Results

```
#plotting health loss
library(ggplot2)
g <- ggplot(data = Top10_events_people, aes(x = reorder(EVTYPE, PEOPLE_LOSS), y = PEOPLE_LOSS))
```

```r
g <- g + geom_bar(stat = "identity", colour = "black")
g <- g + labs(title = "Total people loss in USA by weather events in 1996-2011")
g <- g + theme(plot.title = element_text(hjust = 0.5))
g <- g + labs(y = "Number of fatalities and injuries", x = "Event Type")
g <- g + coord_flip()
print(g)
```

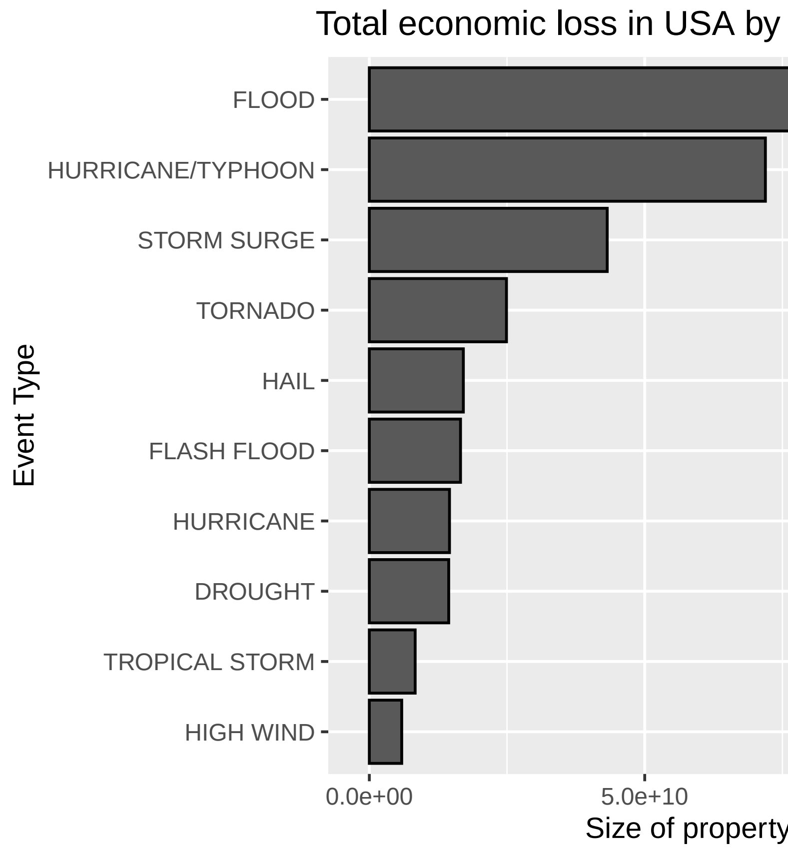
Analyzing population health impact on the graph one can conclude that TORNADOS, EXCESSIVE HEAT and FLOOD are the main contributors to deaths and injuries out of all event types



of weather events.

```r
#plotting economic loss
g <- ggplot(data = Top10_events_economy, aes(x = reorder(EVTYPE, ECONOMIC_LOSS), y = ECONOMIC_LOSS))
g <- g + geom_bar(stat = "identity", colour = "black")
g <- g + labs(title = "Total economic loss in USA by weather events in 1996-2011")
g <- g + theme(plot.title = element_text(hjust = 0.5))
g <- g + labs(y = "Size of property and crop loss", x = "Event Type")
g <- g + coord_flip()
print(g)
```

Analyzing economic impact on the graph one can conclude that FLOOD, HURRICANE/TYPHOON and STORM SURGE are the main contributors to severe economic conse-



quences out of all event types of weather events.