Storm impact on the US population health and economy

Synopsis

In the current research the main goal was to apply reproducible research knowledge into the storms dataset. There were 2 main questions: which storm events affect the most on the public health and witch storm events affect the US economy the most. Results showed, that there are 15 which events, that cover more that 80% of all the impact on the public health and the US economy

Introduction

Before the data processing, first, specify the global options to show all the code and results

```
library(knitr)
library(dplyr)
```

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

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

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

```
opts_chunk$set(echo = TRUE, results = TRUE)
```

Data processing

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size

```
fileUrl <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
fileDest <- ("storm_data.csv.bz2")
if(!file.exists(fileDest)){
  download.file(fileUrl, fileDest)
}
storm <- read.csv("storm_data.csv.bz2")
```

Data manipulation

Most harmful storm events on the population health

Before analyzing, count the unique values of event types (EVTYPEs)

```
length(unique(storm$EVTYPE))
```

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

Almost a thousand, which is many. Let's count total injuries and fatal cases per each type and sort the result by descending order or fatal cases.

```
by_type <- storm %>%
  group_by(EVTYPE) %>%
  summarise(sum(INJURIES), sum(FATALITIES)) %>%
  rename(fatal = 'sum(FATALITIES)', injury = 'sum(INJURIES)') %>%
  filter(fatal != 0, injury !=0) %>%
  arrange(desc(fatal, injury))
```

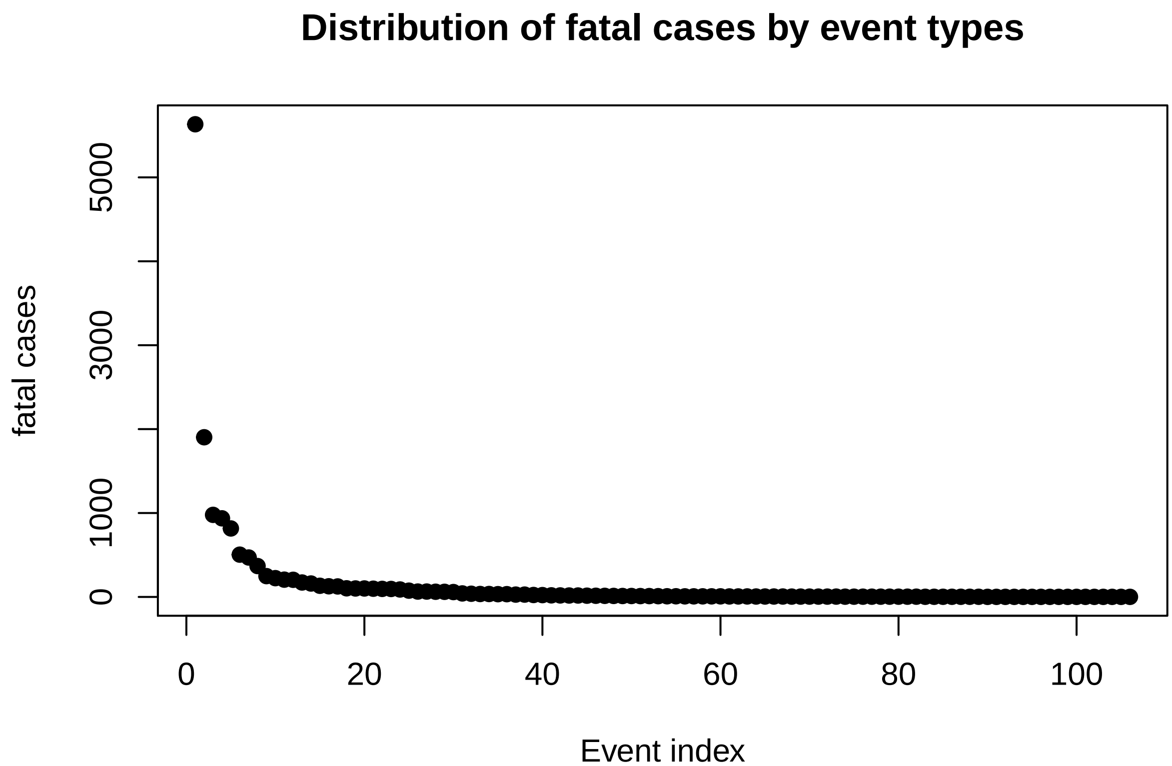
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(by_type)
```

```
## # A tibble: 6 x 3
##   EVTYPE         injury fatal
##   <chr>           <dbl> <dbl>
## 1 TORNADO         91346  5633
## 2 EXCESSIVE HEAT   6525  1903
## 3 FLASH FLOOD      1777   978
## 4 HEAT             2100   937
## 5 LIGHTNING        5230   816
## 6 TSTM WIND        6957   504
```

To better understand the distribution of fatal cases, we can plot them.

```
plot(by_type$fatal, pch = 19, ylab = "fatal cases", xlab= "Event index",
     main = "Distribution of fatal cases by event types")
```



From the plot, we see there are several event types, that has the most of fatal cases. To get the list of the events, that have the most impact, I will be using 80% rule, keep those types, that in total produce 80% of all the fatal cases.

Also, because only several events cover most of the distribution, there is no need to wrangle with other names of events.

```
topfatal <- by_type %>%
  mutate(cumsum.prop = cumsum(fatal)/sum(fatal)) %>%
  filter(cumsum.prop <= 0.8)
topfatal
```

```
## # A tibble: 9 x 4
##   EVTYPE         injury fatal cumsum.prop
##   <chr>           <dbl> <dbl>       <dbl>
## 1 TORNADO         91346  5633       0.377
## 2 EXCESSIVE HEAT   6525  1903       0.504
## 3 FLASH FLOOD      1777   978       0.570
## 4 HEAT             2100   937       0.633
## 5 LIGHTNING        5230   816       0.687
## 6 TSTM WIND        6957   504       0.721
## 7 FLOOD            6789   470       0.752
## 8 RIP CURRENT       232   368       0.777
## 9 HIGH WIND        1137   248       0.794
```

Thus, only 9 event types fit into the criteria and become most harmful on the population health, namely: tornado, excessive heat, flash flood, heat, lightning, tstm wind, flood, rip current, high wind.

Most harmful storm events on the US economy

In this section we will be calculating the impact on the economy by looking at the property and crop damage.

First we need to prepare the data to be proceeded.

```
economydmg <- select(storm, EVTYPE, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP) %>%
  filter(PROPDMG != 0 | CROPDMG != 0)
head(economydmg)
```

```
##    EVTYPE PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 TORNADO    25.0          K       0
## 2 TORNADO     2.5          K       0
## 3 TORNADO    25.0          K       0
## 4 TORNADO     2.5          K       0
## 5 TORNADO     2.5          K       0
## 6 TORNADO     2.5          K       0
```

```
unique(economydmg$PROPDMGEXP, economydmg)
```

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

Because the data has an exponent value, we need to create 2 new features, that multiply initial number into the exponent, where B or b = Billion, M or m = Million, K or k = Thousand, H or h = Hundred

Calculating property damage

```
propdmgcomb <- c()
for (i in 1:nrow(economydmg)){
  if(economydmg$PROPDMGEXP[i] == "K"){
    propdmgcomb[i] <- economydmg$PROPDMG[i] * 1000
  } else if(economydmg$PROPDMGEXP[i] == "m" | economydmg$PROPDMGEXP[i] == "M"){
    propdmgcomb[i] <- economydmg$PROPDMG[i] * 1000000
  } else if(economydmg$PROPDMG[i] == "B"){
    propdmgcomb[i] <- economydmg$PROPDMG[i] * 1000000000
  } else if(economydmg$PROPDMG[i] == "h" | economydmg$PROPDMG[i] == "H"){
    propdmgcomb[i] <- economydmg$PROPDMG[i] * 100
  } else {
    propdmgcomb[i] <- economydmg$PROPDMG[i]
  }
}
summary(propdmgcomb)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##         0      2500     10000    618161     40000 929000000
```

Calculating crop damage

```
cropdmgcomb <- c()
for (i in 1:nrow(economydmg)){
  if(economydmg$CROPDMGEXP[i] == "K"){
    cropdmgcomb[i] <- economydmg$CROPDMG[i] * 1000
```

```
  } else if(economydmg$CROPDMGEXP[i] == "m" | economydmg$CROPDMGEXP[i] == "M"){
    cropdmgcomb[i] <- economydmg$CROPDMG[i] * 1000000
  } else if(economydmg$CROPDMGEXP[i] == "B"){
    cropdmgcomb[i] <- economydmg$CROPDMG[i] * 1000000000
  } else if(economydmg$CROPDMGEXP[i] == "h" | economydmg$CROPDMGEXP[i] == "H"){
    cropdmgcomb[i] <- economydmg$CROPDMG[i] * 100
  } else {
    cropdmgcomb[i] <- economydmg$CROPDMG[i]
  }
}
summary(cropdmgcomb)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 0.000e+00 0.000e+00 2.004e+05 0.000e+00 5.000e+09
```

Now combine resulted vectors with an *economydmg* dataset

```
economydmg <- cbind(economydmg, propdmgcomb, cropdmgcomb)
names(economydmg)
```

```
## [1] "EVTYPE"      "PROPDMG"     "PROPDMGEXP"  "CROPDMG"     "CROPDMGEXP"
## [6] "propdmgcomb" "cropdmgcomb"
```

Finally, calculate which storm type affect more on the economy by property damage and crop damage, and create a new feature, that combines them together.

```
totdamage <- economydmg %>%
  group_by(EVTYPE) %>%
  summarise(sum(propdmgcomb), sum(cropdmgcomb)) %>%
  rename(propdmgtot = 'sum(propdmgcomb)', cropdmgtot = 'sum(cropdmgcomb)') %>%
  mutate(totaldmg = propdmgtot + cropdmgtot) %>%
  arrange(-totaldmg)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(totdamage)
```

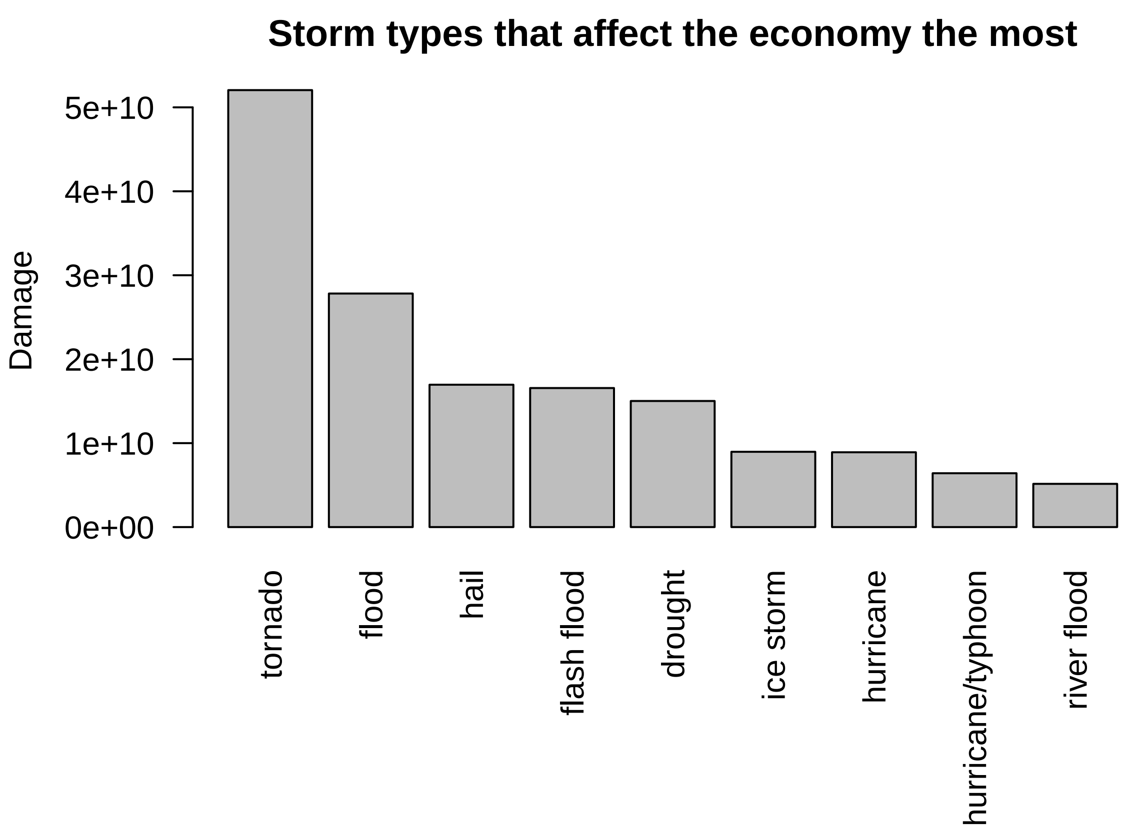
```
## # A tibble: 6 x 4
##   EVTYPE        propdmgtot  cropdmgtot     totaldmg
##   <chr>              <dbl>       <dbl>        <dbl>
## 1 TORNADO      51637160784    414953270 52052114054
## 2 FLOOD        22157709930.  5661968450 27819678380.
## 3 HAIL         13932267050.  3025537890 16957804940.
## 4 FLASH FLOOD 15140812068.  1421317100 16562129168.
## 5 DROUGHT       1046106000 13972566000 15018672000
## 6 ICE STORM     3944927860  5022113500  8967041360
```

Now use the Pareto 80% rule to get a list of the most influential storm types

```
pareto_economy <- totdamage %>%
  mutate(dmg_cumsumprop = cumsum(totaldmg)/sum(totaldmg)) %>%
  filter(dmg_cumsumprop <= 0.8)
```

And make a plot with the final list

```
par(mar=c(8,5.5,3,2))
barplot(pareto_economy$totaldmg, names.arg = tolower(pareto_economy$EVTYPE)
        ,las = 2, main = "Storm types that affect the economy the most")
title(ylab="Damage", mgp=c(4,1,0))
```



Results

In the research we have found, that 9 storm types affect nearly 80% of public health. And similarly, 15 types affect 80 economy damage. In the table you can see final results

```
combine_result <- data.frame(rank = seq(1:9), Public.Health = topfatal$EVTYPE,
           Economy.Damage = pareto_economy$EVTYPE)
inter <- intersect(combine_result$Public.Health, combine_result$Economy.Damage)
length(unique(c(topfatal$EVTYPE, pareto_economy$EVTYPE)))
```

```
## [1] 15
```

```
combine_result
```

```
##   rank  Public.Health    Economy.Damage
## 1    1        TORNADO           TORNADO
## 2    2 EXCESSIVE HEAT             FLOOD
## 3    3    FLASH FLOOD              HAIL
## 4    4           HEAT       FLASH FLOOD
## 5    5      LIGHTNING           DROUGHT
## 6    6      TSTM WIND         ICE STORM
## 7    7          FLOOD         HURRICANE
## 8    8    RIP CURRENT HURRICANE/TYPHOON
## 9    9      HIGH WIND       RIVER FLOOD
```

There are 15 events, that cover more than 80% of all the public health and US economy together. But only 3 of them appear in the top 9 lists, that are: TORNADO, FLASH FLOOD, FLOOD