Storm data analysis in the U.S.

Synopsis

This project involves exploring the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. The events in the database start in the year 1950 and end in November 2011.

Original database: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2

The basic goal of the analisys is to find out which types of events are most harmful with respect to population health and which types of events have the greatest economic consequences across the United States.

Data Processing

Install packages and include all necessary libraries:

```
library(R.utils)
```

```
## Loading required package: R.oo
## Loading required package: R.methodsS3
## R.methodsS3 v1.7.0 (2015-02-19) successfully loaded. See ?R.methodsS3 for help.
## R.oo v1.19.0 (2015-02-27) successfully loaded. See ?R.oo for help.
##
## Attaching package: 'R.oo'
##
## The following objects are masked from 'package:methods':
##
##     getClasses, getMethods
##
## The following objects are masked from 'package:base':
##
##     attach, detach, gc, load, save
##
## R.utils v2.1.0 (2015-05-27) successfully loaded. See ?R.utils for help.
##
## Attaching package: 'R.utils'
##
## The following object is masked from 'package:utils':
##
##     timestamp
##
## The following objects are masked from 'package:base':
##
##     cat, commandArgs, getOption, inherits, isOpen, parse, warnings
```

```
library(plyr)
library(ggplot2)
library(gridExtra)
```

Import data

1. Download file which contains data:

```
download.file("http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2", "~/storm.csv.bz
```

2. Unzip file and read data into variable called 'data':

```
bunzip2("~/storm.csv.bz2", "~/storm.csv", remove = FALSE)
data <- read.csv("~/storm.csv")
```

3. Check dataset parameters:

```
format(object.size(data), units = "Mb")
```

```
## [1] "429335808"
```

```
head(data,3)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1 4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1 4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3       1 2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
## 3 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14.0   100 3   0          0
## 2         NA         0                       2.0   150 2   0          0
## 3         NA         0                       0.1   123 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1       15    25.0          K       0
## 2        0     2.5          K       0
## 3        2    25.0          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
## 3     3340      8742          0          0              3
```

```
names(data)
```

```
##  [1] "STATE__"    "BGN_DATE"   "BGN_TIME"   "TIME_ZONE"  "COUNTY"
##  [6] "COUNTYNAME" "STATE"      "EVTYPE"     "BGN_RANGE"  "BGN_AZI"
## [11] "BGN_LOCATI" "END_DATE"   "END_TIME"   "COUNTY_END" "COUNTYENDN"
## [16] "END_RANGE"  "END_AZI"    "END_LOCATI" "LENGTH"     "WIDTH"
## [21] "F"          "MAG"        "FATALITIES" "INJURIES"   "PROPDMG"
## [26] "PROPDMGEXP" "CROPDMG"    "CROPDMGEXP" "WFO"        "STATEOFFIC"
## [31] "ZONENAMES"  "LATITUDE"   "LONGITUDE"  "LATITUDE_E" "LONGITUDE_"
## [36] "REMARKS"    "REFNUM"
```

```
dim(data)
```

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

Find top 10 most harmful events to population health

In order to find which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health the following steps should be done:

1. Subset original data set to the column: EVTYPE,FATALITIES,INJURIES
2. Calculate statistical information
3. Choose top 10 most harmful fatalities and injuries

```
health <- subset (data, select = c(EVTYPE,FATALITIES,INJURIES))
stat <- ddply(health, .(EVTYPE), summarize,fatalities = sum(FATALITIES),injuries = sum(INJURIES))
fatalities <- head(stat[order(stat$fatalities, decreasing = TRUE),c(1,2)],10)
injuries <- head(stat[order(stat$injuries, decreasing = TRUE),c(1,3)],10)
```

Find types of events which have the greatest economic consequences

In order to find which types of events have the greatest economic consequences the following steps should be done:

1. Subset original data set to the column: EVTYPE,PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGEXP, where PROPDMGEXP and CROPDMGEXP contains a multiplier (Hundred (H), Thousand (K), Million (M), Billion (B)) for PROPDMG (property damage) and CROPDMG
2. Convert characters to numbers by calling convert()
3. Calculate total damage for each parameter in US dollars and statistics
4. Sort resuls in decreasing order

```
economic <- subset (data, select = c(EVTYPE,PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGEXP))
convert <- function (dmexp) {
  dmexp <- toupper(dmexp)
  dmexp <- ifelse(dmexp == ''|dmexp == '0', 1, dmexp)
  dmexp <- ifelse(dmexp == '1', 10, dmexp)
  dmexp <- ifelse(dmexp == 'H'|dmexp == '2', 100, dmexp)
  dmexp <- ifelse(dmexp == 'K'|dmexp == '3', 1e+03, dmexp)
  dmexp <- ifelse(dmexp == '4', 1e+04, dmexp)
  dmexp <- ifelse(dmexp == '5', 1e+05, dmexp)
  dmexp <- ifelse(dmexp == 'M'|dmexp == '6', 1e+06, dmexp)
  dmexp <- ifelse(dmexp == '7', 1e+07, dmexp)
  dmexp <- ifelse(dmexp == '8', 1e+08, dmexp)
  dmexp <- ifelse(dmexp == 'B',1e+09, dmexp)
  dmexp <- ifelse(dmexp == '?'|dmexp == '+'|dmexp == '-', 0, dmexp)
  return (as.numeric(dmexp))
}
economic$prop <- convert(economic$PROPDMGEXP)*economic$PROPDMG
```

```
economic$crop <- convert(economic$CROPDMGEXP)*economic$CROPDMG
economstat <- ddply(economic, .(EVTYPE), summarize,prop = sum(prop)/1000000,crop = sum(crop)/1000000)
prop <- economstat[order(economstat$prop, decreasing = TRUE),c(1,2)]
crop <- economstat[order(economstat$crop, decreasing = TRUE),c(1,3)]
```

Results

Check results for each parameter:

```
head(fatalities)
```

```
##             EVTYPE fatalities
## 830        TORNADO       5633
## 123 EXCESSIVE HEAT       1903
## 147    FLASH FLOOD        978
## 269           HEAT        937
## 452      LIGHTNING        816
## 854      TSTM WIND        504
```

```
head(injuries)
```

```
##             EVTYPE injuries
## 830        TORNADO    91346
## 854      TSTM WIND     6957
## 164          FLOOD     6789
## 123 EXCESSIVE HEAT     6525
## 452      LIGHTNING     5230
## 269           HEAT     2100
```

```
head(prop)
```

```
##                EVTYPE      prop
## 164             FLOOD 144657.71
## 406 HURRICANE/TYPHOON  69305.84
## 830           TORNADO  56947.38
## 666       STORM SURGE  43323.54
## 147       FLASH FLOOD  16822.68
## 238              HAIL  15735.27
```

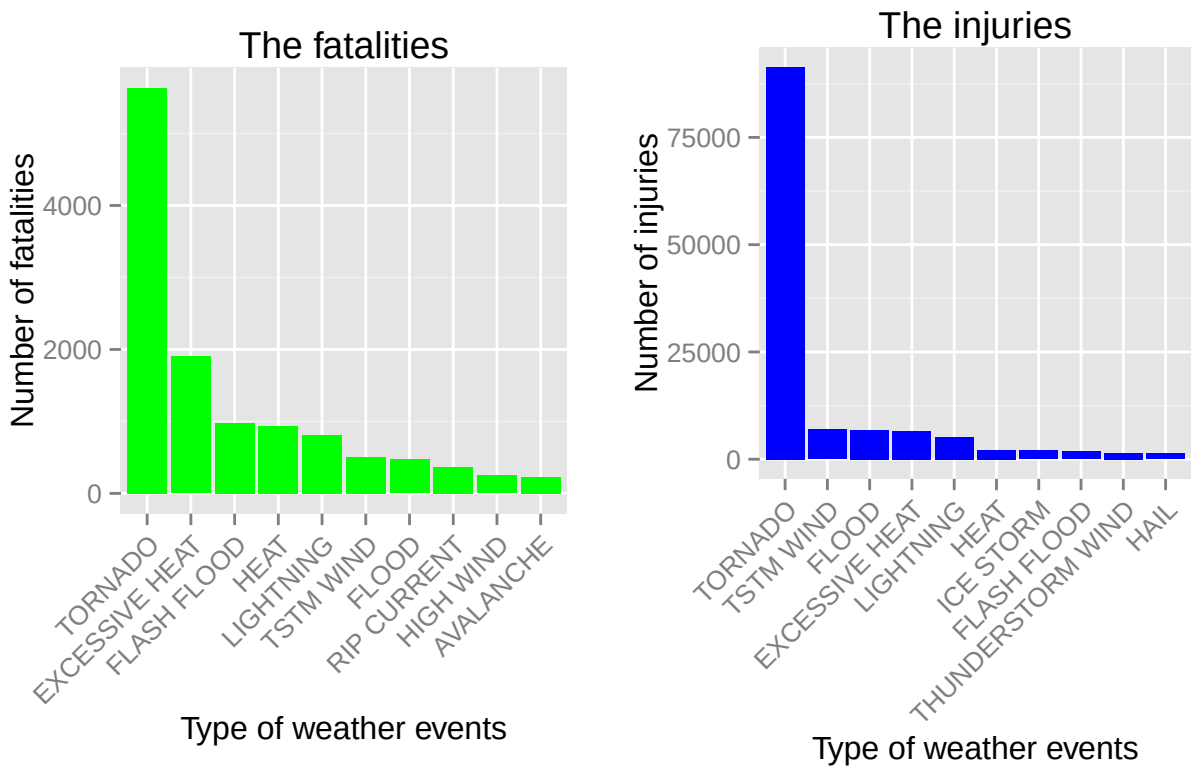
```
head(crop)
```

```
##          EVTYPE      crop
## 88      DROUGHT 13972.566
## 164       FLOOD  5661.968
## 586 RIVER FLOOD  5029.459
## 424   ICE STORM  5022.114
## 238        HAIL  3025.955
## 397   HURRICANE  2741.910
```

Data vizualization

Figure 1: Top 10 most harmful events to population health in the period from 1950 to 2011 in the U.S.

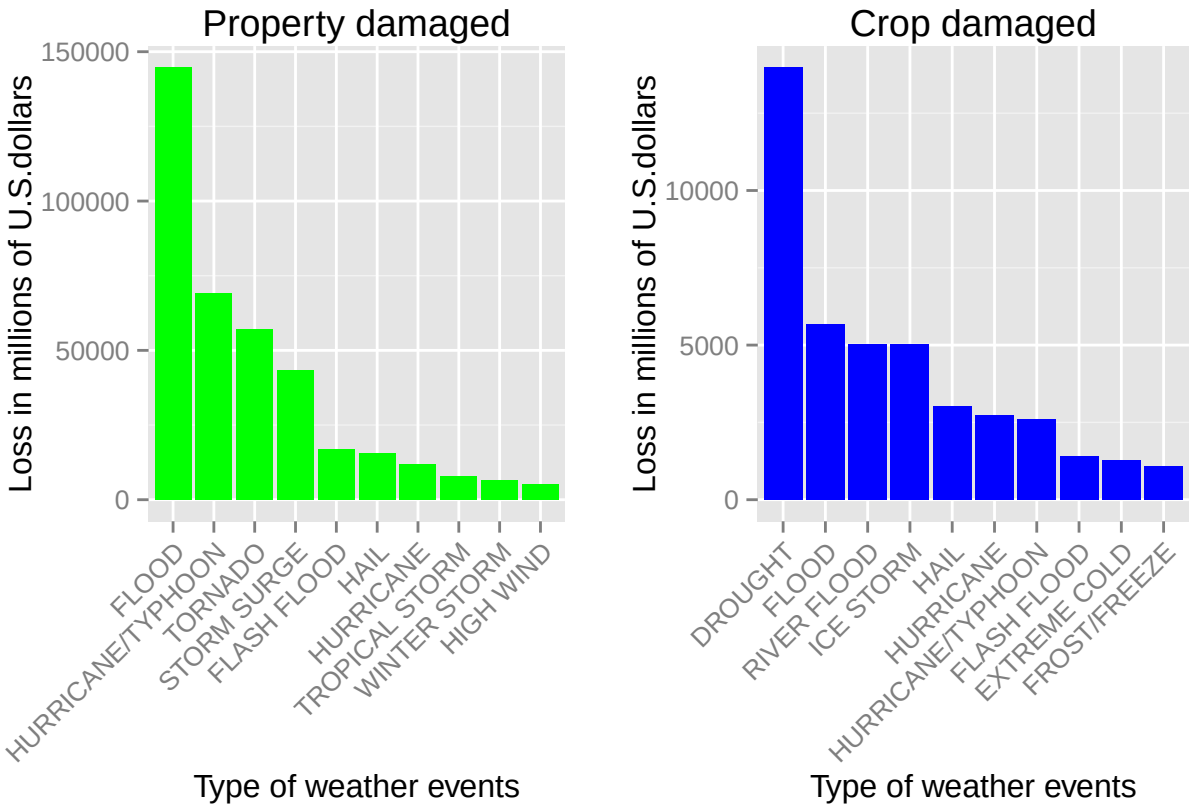
```
#Plots for health stat
fplot <- ggplot(fatalities, aes(x=reorder(EVTYPE, -fatalities), y=fatalities)) +
  geom_bar(stat="identity", fill="green") +
  labs(title="The fatalities", x="Type of weather events", y="Number of fatalities") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), aspect.ratio = 1)
iplot <- ggplot(injuries, aes(x=reorder(EVTYPE, -injuries), y=injuries)) +
  geom_bar(stat="identity", fill="blue") +
  labs(title="The injuries", x="Type of weather events", y="Number of injuries") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), aspect.ratio = 1)
grid.arrange(fplot, iplot, ncol=2)
```



According to the analysis Tornado can be considered as the most harmful event.

Figure 2: Economic impact (in millins dollars) by weather events in U.S. in the period from 1950 to 2011

```
#Plots for economic stat
pplot <- ggplot(head(prop,10), aes(x=reorder(EVTYPE, -prop), y=prop)) +
  geom_bar(stat="identity", fill="green") +
  labs(title = "Property damaged", x="Type of weather events", y="Loss in millions of U.S.dollars") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
cplot <- ggplot(head(crop,10), aes(x=reorder(EVTYPE, -crop), y=crop)) +
  geom_bar(stat="identity", fill="blue") +
  labs(title = "Crop damaged", x="Type of weather events", y="Loss in millions of U.S.dollars") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
grid.arrange(pplot, cplot, ncol=2)
```



In terms of property damage Flood has the biggest impact, in terms of crop damage Drought has the biggest impact.