Reproducible Research - Course project 2

Synopsis

In this report, I aim to describe several weather events in the United States between the years 1950 and 2011. The data are taken from the U.S. National Oceanic and Atmospheric Administration's (NOAA) and include weather events informations, like when and where they occur, estimates of any fatalities, injuries and property damage. The goal is to analyze the data and answer the following questions:

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?

After loading and processing the data, I came to the following conclusions:

1- Tornado is the most harmful event in terms of human fatalites and injuries;

2- Floods have the greatest economic consequences.

The database documentation can be found here and here. The following sections describe the steps taken to analyze the data.

Data Processing

The data are downloaded from NOAA Storm Database, if necessary, and read to R.

```
if (!file.exists('./storm_data')) { dir.create('./storm_data') }
if (!file.exists("storm_data/repdata_data_StormData.csv.bz2")) {
  download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStor
mData.csv.bz2", destfile="storm_data/repdata_data_StormData.csv.bz2", mode
= "wb",  method = "curl")
}
if (!exists('storm_data')) {
  storm_data <- read.csv("storm_data/repdata_data_StormData.csv.bz2")
}
```

The required packages are plyr and ggplot2:

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

A brief summary of the data.

```
summary(storm_data)
```

```
##      STATE__                  BGN_DATE             BGN_TIME
##  Min.   : 1.0   5/25/2011 0:00:00:  1202   12:00:00 AM: 10163
##  1st Qu.:19.0   4/27/2011 0:00:00:  1193   06:00:00 PM:  7350
##  Median :30.0   6/9/2011 0:00:00 :  1030   04:00:00 PM:  7261
##  Mean   :31.2   5/30/2004 0:00:00:  1016   05:00:00 PM:  6891
##  3rd Qu.:45.0   4/4/2011 0:00:00 :  1009   12:00:00 PM:  6703
##  Max.   :95.0   4/2/2006 0:00:00 :   981   03:00:00 PM:  6700
##                 (Other)          :895866   (Other)    :857229
##    TIME_ZONE          COUNTY           COUNTYNAME         STATE
##  CST    :547493   Min.   :  0.0   JEFFERSON :  7840   TX     : 83728
##  EST    :245558   1st Qu.: 31.0   WASHINGTON:  7603   KS     : 53440
##  MST    : 68390   Median : 75.0   JACKSON   :  6660   OK     : 46802
##  PST    : 28302   Mean   :100.6   FRANKLIN  :  6256   MO     : 35648
##  AST    :  6360   3rd Qu.:131.0   LINCOLN   :  5937   IA     : 31069
##  HST    :  2563   Max.   :873.0   MADISON   :  5632   NE     : 30271
##  (Other):  3631                   (Other)   :862369   (Other):621339
##                EVTYPE         BGN_RANGE           BGN_AZI
##  HAIL             :288661   Min.   :   0.000          :547332
##  TSTM WIND        :219940   1st Qu.:   0.000   N      : 86752
##  THUNDERSTORM WIND: 82563   Median :   0.000   W      : 38446
##  TORNADO          : 60652   Mean   :   1.484   S      : 37558
##  FLASH FLOOD      : 54277   3rd Qu.:   1.000   E      : 33178
##  FLOOD            : 25326   Max.   :3749.000   NW     : 24041
##  (Other)          :170878                      (Other):134990
##          BGN_LOCATI                  END_DATE             END_TIME
##                :287743                     :243411              :238978
##  COUNTYWIDE    : 19680   4/27/2011 0:00:00:  1214   06:00:00 PM:  9802
##  Countywide    :   993   5/25/2011 0:00:00:  1196   05:00:00 PM:  8314
##  SPRINGFIELD   :   843   6/9/2011 0:00:00 :  1021   04:00:00 PM:  8104
##  SOUTH PORTION:   810   4/4/2011 0:00:00 :  1007   12:00:00 PM:  7483
##  NORTH PORTION:   784   5/30/2004 0:00:00:   998   11:59:00 PM:  7184
##  (Other)       :591444   (Other)          :653450   (Other)    :622432
##    COUNTY_END COUNTYENDN       END_RANGE           END_AZI
##  Min.   :0    Mode:logical   Min.   :  0.0000          :724837
##  1st Qu.:0    NA's:902297    1st Qu.:  0.0000   N      : 28082
##  Median :0                   Median :  0.0000   S      : 22510
##  Mean   :0                   Mean   :  0.9862   W      : 20119
##  3rd Qu.:0                   3rd Qu.:  0.0000   E      : 20047
##  Max.   :0                   Max.   :925.0000   NE     : 14606
##                                                 (Other): 72096
##            END_LOCATI          LENGTH              WIDTH
##                 :499225   Min.   :   0.0000   Min.   :   0.000
##  COUNTYWIDE     : 19731   1st Qu.:   0.0000   1st Qu.:   0.000
##  SOUTH PORTION  :   833   Median :   0.0000   Median :   0.000
##  NORTH PORTION  :   780   Mean   :   0.2301   Mean   :   7.503
##  CENTRAL PORTION:   617   3rd Qu.:   0.0000   3rd Qu.:   0.000
##  SPRINGFIELD    :   575   Max.   :2315.0000   Max.   :4400.000
##  (Other)        :380536
##         F               MAG            FATALITIES          INJURIE
S
##  Min.   :0.0      Min.   :    0.0   Min.   :  0.0000   Min.   :   0.000
0
```

```
##  1st Qu.:0.0      1st Qu.:    0.0   1st Qu.:  0.0000   1st Qu.:   0.000
0
##  Median :1.0      Median :   50.0   Median :  0.0000   Median :   0.000
0
##  Mean   :0.9      Mean   :   46.9   Mean   :  0.0168   Mean   :   0.155
7
##  3rd Qu.:1.0      3rd Qu.:   75.0   3rd Qu.:  0.0000   3rd Qu.:   0.000
0
##  Max.   :5.0      Max.   :22000.0   Max.   :583.0000   Max.   :1700.000
0
##  NA's   :84356
3
##     PROPDMG           PROPDMGEXP        CROPDMG           CROPDMGEXP
##  Min.   :   0.00          :465934   Min.   :  0.000          :618413
##  1st Qu.:   0.00   K      :424665   1st Qu.:  0.000   K      :281832
##  Median :   0.00   M      : 11330   Median :  0.000   M      :  1994
##  Mean   :  12.06   0      :   216   Mean   :  1.527   k      :    21
##  3rd Qu.:   0.50   B      :    40   3rd Qu.:  0.000   0      :    19
##  Max.   :5000.00   5      :    28   Max.   :990.000   B      :     9
##                    (Other):    84                     (Other):     9
##       WFO                                       STATEOFFIC
##         :142069                                       :248769
##  OUN    : 17393   TEXAS, North                       : 12193
##  JAN    : 13889   ARKANSAS, Central and North Central: 11738
##  LWX    : 13174   IOWA, Central                      : 11345
##  PHI    : 12551   KANSAS, Southwest                  : 11212
##  TSA    : 12483   GEORGIA, North and Central         : 11120
##  (Other):690738   (Other)                            :595920
#
#

                                                ZONENAMES
#
#

                                                     :594029
#
#

                                                     :205988
##  GREATER RENO / CARSON CITY / M - GREATER RENO / CARSON CITY /
M

                                                                 :   639
##  GREATER LAKE TAHOE AREA - GREATER LAKE TAHOE ARE
A

:   592
##  JEFFERSON - JEFFERSO
N

                            :   303
##  MADISON - MADISO
```

```
N

                                  :   302
##  (Othe
r)

                                           :100444
##     LATITUDE      LONGITUDE        LATITUDE_E     LONGITUDE_
##  Min.   :   0   Min.   :-14451   Min.   :   0   Min.   :-14455
##  1st Qu.:2802   1st Qu.:  7247   1st Qu.:   0   1st Qu.:     0
##  Median :3540   Median :  8707   Median :   0   Median :     0
##  Mean   :2875   Mean   :  6940   Mean   :1452   Mean   :  3509
##  3rd Qu.:4019   3rd Qu.:  9605   3rd Qu.:3549   3rd Qu.:  8735
##  Max.   :9706   Max.   : 17124   Max.   :9706   Max.   :106220
##  NA's   :47                      NA's   :40
##                                             REMARKS           REFNUM
##                                                 :287433   Min.   :     1
##                                                 : 24013   1st Qu.:225575
##  Trees down.\n                                  :  1110   Median :451149
##  Several trees were blown down.\n               :   568   Mean   :451149
##  Trees were downed.\n                           :   446   3rd Qu.:676723
##  Large trees and power lines were blown down.\n:   432   Max.   :902297
##  (Other)                                        :588295
```

The selected variables, which are more likely to address the initial questions are: fatalities, injuries, property damage and crop damage.

```
summary(storm_data$FATALITIES)
```

```
##      Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0000   0.0000   0.0168   0.0000 583.0000
```

```
summary(storm_data$INJURIES)
```

```
##       Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##     0.0000    0.0000    0.0000    0.1557    0.0000 1700.0000
```

```
summary(storm_data$PROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   12.06    0.50 5000.00
```

```
summary(storm_data$CROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.527   0.000 990.000
```

According to the documentation (page 12), the PROPDMG and CROPDMG variables are encoded by PROPDMGEXP and CROPDMGEXP variables, respectively. They represent magnitude values, including “H” for hundreds, “K” for thousands, “M”" for millions and “B” for billions. In order to decode PROPDMG and CROPDMG, I created new numeric variables.

```
storm_data$PROPMULT <- 1
storm_data$PROPMULT[storm_data$PROPDMGEXP =="H"] <- 100
storm_data$PROPMULT[storm_data$PROPDMGEXP =="K"] <- 1000
storm_data$PROPMULT[storm_data$PROPDMGEXP =="M"] <- 1000000
storm_data$PROPMULT[storm_data$PROPDMGEXP =="B"] <- 1000000000

storm_data$CROPMULT <- 1
storm_data$CROPMULT[storm_data$CROPDMGEXP =="H"] <- 100
storm_data$CROPMULT[storm_data$CROPDMGEXP =="K"] <- 1000
storm_data$CROPMULT[storm_data$CROPDMGEXP =="M"] <- 1000000
storm_data$CROPMULT[storm_data$CROPDMGEXP =="B"] <- 1000000000
```

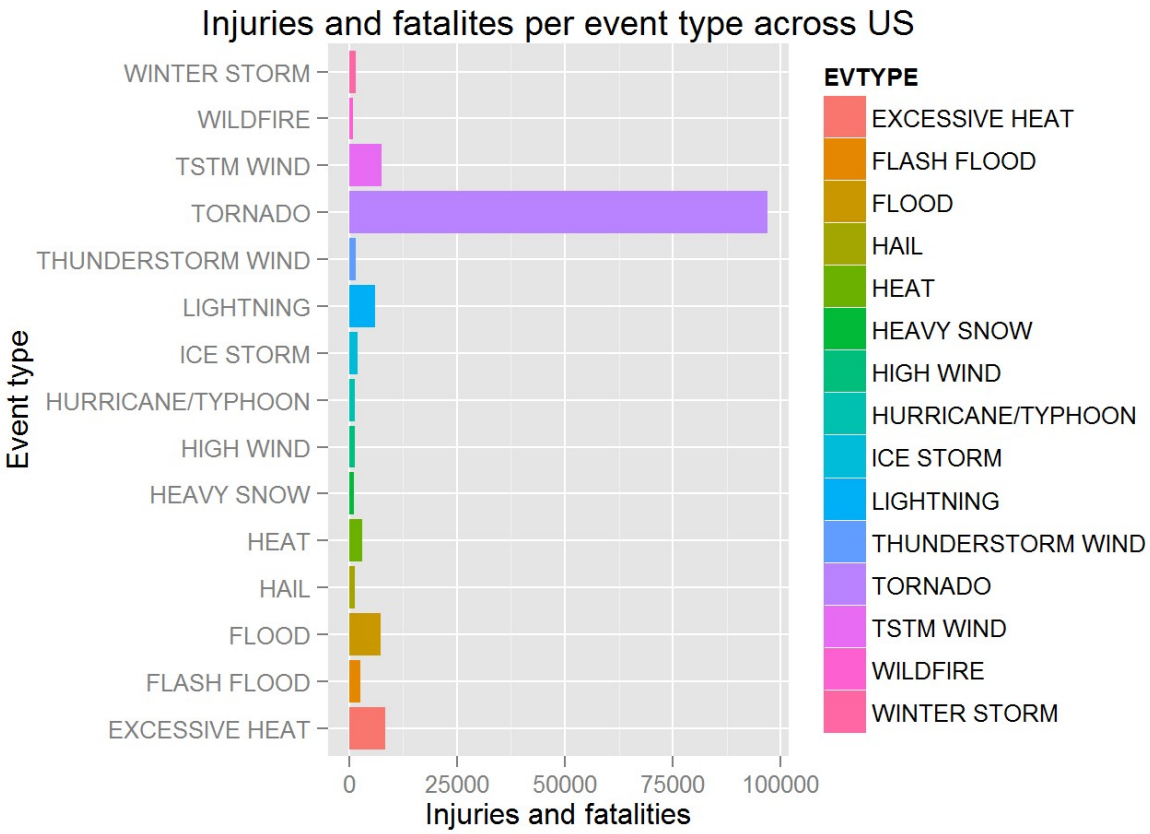
Results

```
aggregate_data <- ddply(.data = storm_data, .variables = .(EVTYPE), fataliti
es = sum(FATALITIES), injuries = sum(INJURIES), property_damage = sum(PROPDM
G * PROPMULT), crop_damage = sum(CROPDMG * CROPMULT), summarize)

population_data <- arrange(aggregate_data, desc(fatalities + injuries))
damage_data <- arrange(aggregate_data, desc(property_damage + crop_damage))
```

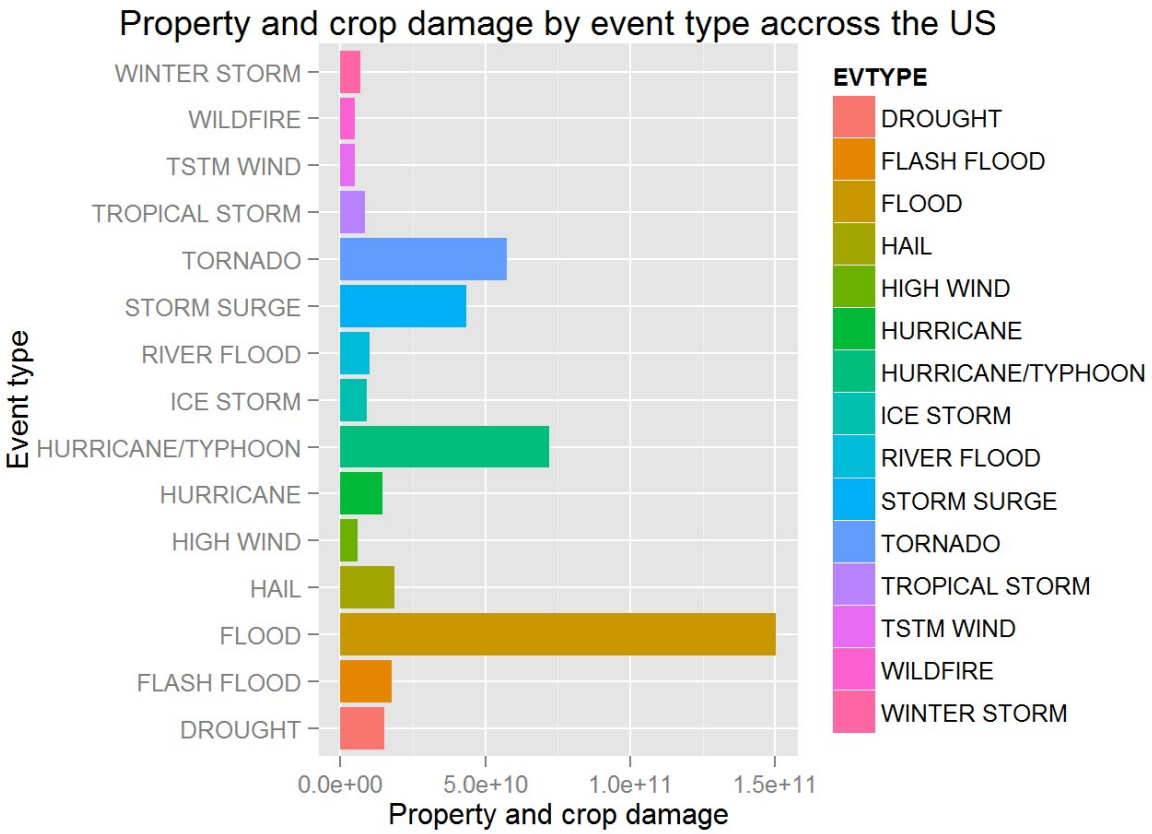
Question 1: Across the United States, which types of events are most harmful with respect to population health?

```
ggplot(data = head(population_data, 15), aes(x = factor(EVTYPE), y = (fatali
ties + injuries), fill = EVTYPE)) + geom_bar(stat="identity") + coord_flip
() + labs(y = "Injuries and fatalities", x = "Event type", title = "Injurie
s and fatalites per event type across US")
```



Question 2: Across the United States, which types of events have the greatest economic consequences?

```
ggplot(data = head(damage_data, 15), aes(x = factor(EVTYPE), y = (property_d
amage + crop_damage), fill = EVTYPE)) + geom_bar(stat="identity") + coord_fl
ip() + labs(y = "Property and crop damage", x = "Event type", title = "Prope
rty and crop damage by event type accross the US")
```



Conclusions

1.Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

Tornadoes are responsible for the largest proportion of both deaths and injuries out of all event types.

2.Across the United States, which types of events have the greatest economic consequences?

Flooding is responsible for the largeset proportion of total economic impact out of all event types.