Reproducible Research: Peer Assessment 2

Events with the most impact on economy and public health in USA

Synopsis

The goal of this report is to explore the US National Oceanic and Atmospheric Administration's (NOAA) storm database and answer two questions about severe weather events, from the time period 1950 to 2011.

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?

The results show that during that timeframe, tornados are most harmful with respect to population health, which have caused 5633 deaths and 91346 injuries.

Tornadoes have cost the most property damages costing over 1.5 billion dollars in economic losses. Excessive wetness have cost the most crop damage costing over 140 million dollars.

Data processing

The data set was downloaded from the course page . Description of the dataset can be found here and here

The main columns we are interested in :

- EVTYPE : Event type (tornado, blizzard, flood,. . . .)
- FATALITIES : people that died during the natural disaster
- INJURIES : people that got injured during the natural disaster
- PROPDMG : property damage
- PROPDMGEXP : property daamge exponent
- CROPDMG : crop damage
- CROPDMGEXP : crop damage exponent

Load required libraries

```r
library(plyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.1.2
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
##
## The following object is masked from 'package:stats':
##
```

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

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.1.2
```

```
library(ggplot2)
library(car)
```

Processing the Data

```
stormdata <- read.csv("repdata-data-StormData.csv.bz2")
```

Dimensions of the dataset

```
dim(stormdata)
```

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

Although there are 37 cols, for our analysis we are only interested in the 7 columns mentioned above.

Here is how the raw data looks like for the columns we are interested

```
head(select(stormdata, EVTYPE,FATALITIES,INJURIES,PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGEXP))
```

```
##    EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 TORNADO          0       15    25.0          K       0
## 2 TORNADO          0        0     2.5          K       0
## 3 TORNADO          0        2    25.0          K       0
## 4 TORNADO          0        2     2.5          K       0
## 5 TORNADO          0        2     2.5          K       0
## 6 TORNADO          0        6     2.5          K       0
```

Info on different event types

```
event_types<-unique(stormdata$EVTYPE)
unique_event<-length(unique(stormdata$EVTYPE))
```

There are 985 events which is a lot to process. Since we are only interested in events that caused the most financial damange and human casualties, we will limit the events by the resepctive damage.

The crop and property damage has the damage broken down by the base and the exponent. Below is the distribution of those symbols:

```r
table(stormdata$PROPDMGEXP)
```

```
##
##              -      ?      +      0      1      2      3      4      5
## 465934       1      8      5    216     25     13      4      4     28
##      6       7      8      B      h      H      K      m      M
##      4       5      1     40      1      6 424665      7  11330
```

To clean it, we follow the below stratergy:

- values such as '-','+','?' are mapped to 0
- For the si exponent or the exponent, we map them to the full numeric base. . . . . . So, k,K,3 becomes 10^3

```r
# make 'k', 'K' map to lowercase
stormdata$PROPDMGEXP <- factor(tolower(stormdata$PROPDMGEXP))
stormdata$CROPDMGEXP <- factor(tolower(stormdata$CROPDMGEXP))

# map the base
stormdata$PROPDMGEXP <- as.numeric(recode(as.character(stormdata$PROPDMGEXP),
    "'0'=1;'1'=10;'2'=10^2;'3'=10^3;'4'=10^4;'5'=10^5;'6'=10^6;'7'=10^7;'8'=10^8;'b'=10^9;'h'=10^2;'k'=
stormdata$CROPDMGEXP <- as.numeric(recode(as.character(stormdata$CROPDMGEXP),
    "'0'=1;'1'=10;'2'=10^2;'3'=10^3;'4'=10^4;'5'=10^5;'6'=10^6;'7'=10^7;'8'=10^8;'b'=10^9;'h'=10^2;'k'=

# calculate value in dollars
stormdata$PROPDMGDOLLAR <- stormdata$PROPDMG * stormdata$PROPDMGEXP
stormdata$CROPDMGDOLLAR <- stormdata$CROPDMG * stormdata$CROPDMGEXP
```

Results

Question 1

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

From the dataset, the columns "FATALITIES" and "INJURIES" which describes human deaths and injuries respectively caused due to the event.

```r
evt_fat_injuries<-ddply(stormdata, .(EVTYPE), function(x) colSums(subset(x, select= c(FATALITIES,INJURI

evt_fat_injuries_by_fat<-arrange(evt_fat_injuries,desc(FATALITIES))%>%  select(EVTYPE,FATALITIES)
evt_fat_injuries_by_inj<-arrange(evt_fat_injuries,desc(INJURIES))%>%  select(EVTYPE,INJURIES)
```

Fatalities sorted by event in descending order

```r
head(evt_fat_injuries_by_fat)
```

```
##             EVTYPE FATALITIES
## 1          TORNADO       5633
## 2 EXCESSIVE HEAT       1903
```

```
## 3    FLASH FLOOD        978
## 4           HEAT        937
## 5      LIGHTNING        816
## 6      TSTM WIND        504
```

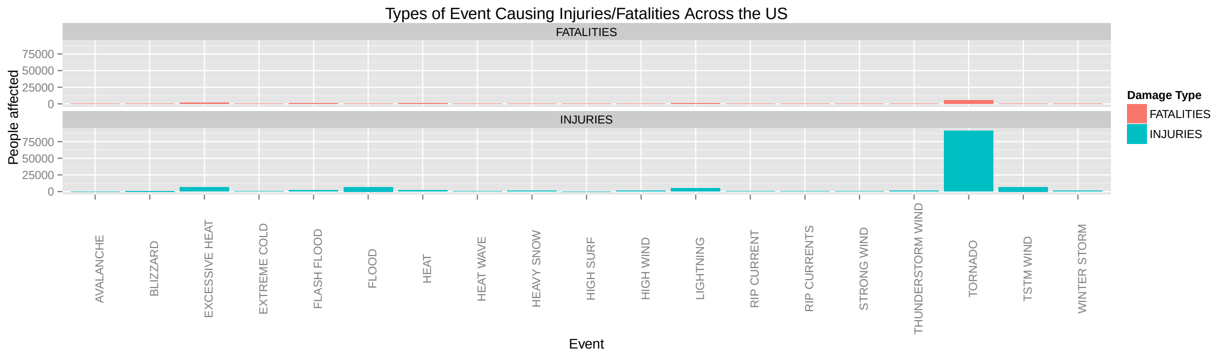
Fatalities sorted by event in descending order

```r
head(evt_fat_injuries_by_inj)
```

```
##          EVTYPE INJURIES
## 1       TORNADO    91346
## 2     TSTM WIND     6957
## 3         FLOOD     6789
## 4 EXCESSIVE HEAT     6525
## 5     LIGHTNING     5230
## 6          HEAT     2100
```

Looking at the data, we see that certain event types cause a lot of fatalities/injuries. So,lets take the minumum casualties we are interested in as greather than 100.

```r
#filter data for casualties greaer than 100
#prepare the data for ggplot
evnts_high_damage<- evt_fat_injuries %>%
  filter(FATALITIES > 100, INJURIES > 100) %>%
  gather(type,people_affected,FATALITIES:INJURIES) %>%
  select(Event=EVTYPE,type,people_affected)


ggplot(evnts_high_damage,aes(x = Event,y = people_affected,fill=type)) +
    geom_bar(stat = "identity") +
    facet_wrap(~type, ncol=1) +
    theme(axis.text.x=element_text(angle = 90)) +
    ggtitle("Types of Event Causing Injuries/Fatalities Across the US") +
    labs(y="People affected",fill = "Damage Type")
```



So, we see that tornados cause the most fatalities(5,633) and injuries (91,346) across the USA from 1950 to November 2011.

It is probably not surpising to see there exists a correlation between the number of injuries and deaths with respect to event.

Question 2

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

Economic consequences can be measured with property damage ("PROPDMGDOLLAR") and crop damage ("CROPDMGDOLLAR")

Aggregate property and crop damage

```
evt_damage<-ddply(stormdata, .(EVTYPE), function(x) colSums(subset(x, select= c(PROPDMGDOLLAR,CROPDMGDOL

evt_damage_for_prop<-arrange(evt_damage,desc(PROPDMGDOLLAR))%>%  select(EVTYPE,PROPDMGDOLLAR)
evt_damage_for_crop<-arrange(evt_damage,desc(CROPDMGDOLLAR))%>%  select(EVTYPE,CROPDMGDOLLAR)
```

Property damage sorted by event in descending order

```
head(evt_damage_for_prop)
```

```
##                         EVTYPE PROPDMGDOLLAR
## 1 TORNADOES, TSTM WIND, HAIL    1600000000
## 2                   WILD FIRES     624100000
## 3                    HAILSTORM     241000000
## 4              HIGH WINDS/COLD     110500000
## 5               River Flooding     106155000
## 6                  MAJOR FLOOD     105000000
```

Crop damage sorted by event in descending order

```
head(evt_damage_for_crop)
```

```
##                        EVTYPE CROPDMGDOLLAR
## 1           EXCESSIVE WETNESS     142000000
## 2   COLD AND WET CONDITIONS      66000000
## 3                 Early Frost      42000000
## 4             Damaging Freeze      34130000
## 5                      Freeze      10500000
## 6 HURRICANE OPAL/HIGH WINDS      10000000
```

Top 20 events that cause maximal crop and property damage

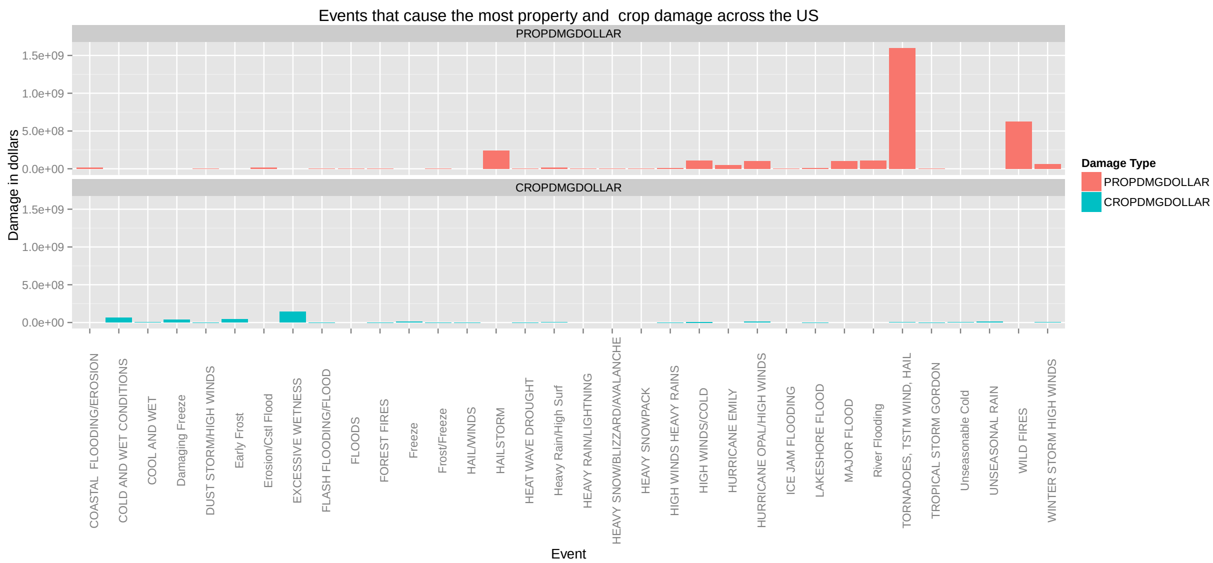
```
top_20_prop_events<-evt_damage_for_prop[1:20,'EVTYPE']
top_20_crop_events<-evt_damage_for_crop[1:20,'EVTYPE']

top_events_damage<-evt_damage[evt_damage$EVTYPE %in% top_20_prop_events | evt_damage$EVTYPE %in% top_20_
```

```
# select top events that cause damage
# prepare the data for ggplot
top_events_damage<- top_events_damage %>%
  gather(type,Damage_in_dollars,PROPDMGDOLLAR:CROPDMGDOLLAR) %>%
  select(Event=EVTYPE,type,Damage_in_dollars)
```

```r
ggplot(top_events_damage,aes(x = Event,y = Damage_in_dollars,fill=type)) +
    geom_bar(stat = "identity") +
    facet_wrap(~type, ncol=1) +
    theme(axis.text.x=element_text(angle = 90)) +
    ggtitle("Events that cause the most property and  crop damage across the US") +
    labs(y="Damage in dollars",fill = "Damage Type")
```

```
## Warning: Removed 9 rows containing missing values (position_stack).
```

```
## Warning: Removed 12 rows containing missing values (position_stack).
```



From the graph, we see that unlike human fatalities and injuries, there is no correlation between property and event damage with respect to event type.

Tornadoes and wild fires have caused the most property damage: 1.600e+09 and 6.241e+08 respectively.

Excessive wetness and "cold and wet conditions" are the biggest source of crop damges: 1.42e+08 and 6.60e+07 respectively.