A Study on Severe Weather Events in USA

Wenhe 'Wayne' Ye

September 24, 2015

Synopsis

1.How to import the raw data?

-Downloading Data

-Caching Data

2.How to clean up the raw data to get tidy data?

-Correcting Typos

-Unit Conversion

3.How to summarize the data?

-Using dplyr package

4.How to present the analysis and result?

-Genrating Tables and Barplots

Data Processing

1.Downloading and importing

This procedure might take a few minutes for downloading and extracting data sets. Raw data stored as *raw_stormdata*

```
rm(list=ls())
if(!dir.exists("./data")){
        dir.create("./data")
}
dataURL<-"https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
if(!file.exists("./data/Stormdata.csv.bz2")){
download.file(dataURL,"./data/Stormdata.csv.bz2")
}
raw_stormdata<-read.csv(bzfile("./data/Stormdata.csv.bz2"))
```

2.Cleaning up and summarizing

We first extract a few columns from the *raw_stormdata*. We are typically interested in variables with regard to either economic loss or life safety.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.2.2
```

```
##
## 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
```

```
my_stormdata<-select(raw_stormdata,EVTYPE,STATE,FATALITIES,INJURIES,PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGE
```

The recorded event type descriptions are messy. We first set them to uppercase and merge the names of similar events into their official categories. This procedure can be done iteratively to ensure the topmost events are recorded correctly. Here in this report, we do not fix all the typos since the some of these are really odd and may not cause a major deviation.

```
my_stormdata<-mutate(my_stormdata,Event_Type=as.factor(toupper(as.character(EVTYPE)))) #set to uppercas
evt_factor_lvl<-levels(my_stormdata$Event_Type)
evt_factor_lvl[which(evt_factor_lvl=="TSTM WIND")]<-"THUNDERSTORM WIND"
evt_factor_lvl[which(evt_factor_lvl=="THUNDERSTORM WINDS")]<-"THUNDERSTORM WIND"
evt_factor_lvl[which(evt_factor_lvl=="WILD/FOREST FIRE")]<-"WILDFIRE"
evt_factor_lvl[which(evt_factor_lvl=="RIP CURRENTS")]<-"RIP CURRENT"
evt_factor_lvl[which(evt_factor_lvl=="HURRICANE")]<-"HURRICANE/TYPHOON"
levels(my_stormdata$Event_Type)<-evt_factor_lvl
```

In order to calculate the right loss data quoted in dollars, we clean up both the *PROPDMGEXP* and *CROPDMGEXP* data and assign a vector *Exp_vector_T* for transform the loss data into trillion dollars. The transformation will be carried out in the analyzing part for Q2.

```
my_stormdata<-mutate(my_stormdata,PropDmgExp=as.factor(toupper(as.character(PROPDMGEXP))))
my_stormdata<-mutate(my_stormdata,CropDmgExp=as.factor(toupper(as.character(CROPDMGEXP))))
prop_fac_lvl<-levels(my_stormdata$PropDmgExp)
prop_fac_lvl[!prop_fac_lvl %in% c("B","M","K","H")]<-"D"
crop_fac_lvl<-levels(my_stormdata$CropDmgExp)
crop_fac_lvl[!crop_fac_lvl %in% c("B","M","K","H")]<-"D"
levels(my_stormdata$PropDmgExp)<- prop_fac_lvl
levels(my_stormdata$CropDmgExp)<- crop_fac_lvl
Exp_vector_T<-c("D"=0.001/1000000000,"H"=0.1/1000000000,"K"=1/1000000000,"M"=1/1000000,"B"=1/1000)
```

(**Q1 Analysis**)We grouped data by event types and calculate the both the total casualties and total fatalities. We saved the top 10 most harmful events in *my_cas_barplot_data*. A barplot will be generated from the resulted data.

```
grouped_stormdata<-group_by(my_stormdata,Event_Type)
sum_cas_stormdata<-summarize(grouped_stormdata,CASUALTIES=sum(FATALITIES+INJURIES),TOTAL_INJURIES=sum(I
sum_cas_stormdata<-arrange(sum_cas_stormdata,desc(CASUALTIES))
sum_cas_stormdata<-mutate(sum_cas_stormdata,CASUALTIES_K=CASUALTIES/1000,TOTAL_FATALITIES_K=TOTAL_FATAL
my_cas_barplot_data<-rbind(sum_cas_stormdata[1:10,]$TOTAL_FATALITIES_K,sum_cas_stormdata[1:10,]$TOTAL_I
names(my_cas_barplot_data)<-as.character(sum_cas_stormdata[1:10,]$Event_Type)
```

(**Q2 Analysis**)We grouped data by event types and calculate the both the total property damage and crop damage. The actual loss in the unit of trillion dollars is calculated by the code below by looking up the vector *Exp_vector_T*. We shortlist the top 10 most dustructive events in *my_eco_barplot_data*. A barplot will be generated from the resulted data.

```
sum_eco_stormdata<-summarize(grouped_stormdata,PROPERTY_LOSS=sum(PROPDMG*Exp_vector_T[PropDmgExp]),CROP_
sum_eco_stormdata<-mutate(sum_eco_stormdata,PROPERTY_LOSS,CROP_LOSS_M=CROP_LOSS*1000,TOTAL_LOSS=PROPERT
sum_eco_stormdata<-arrange(sum_eco_stormdata,desc(TOTAL_LOSS))
my_eco_barplot_data<-rbind(sum_eco_stormdata[1:10,]$PROPERTY_LOSS,sum_eco_stormdata[1:10,]$CROP_LOSS)
names(my_eco_barplot_data)<-as.character(sum_eco_stormdata[1:10,]$Event_Type)
```

Results

(**Anser to Q1**)
Below is the top 10 harmful severe weather events across US.

```
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.2.2
```

```
report_cas_stormdata<-sum_cas_stormdata[1:10,]
xreport_cas<-xtable(select(report_cas_stormdata,Event_Type,CASUALTIES,TOTAL_FATALITIES,TOTAL_INJURIES),
print.xtable(xreport_cas,type = "html")
```

Top 10 harmful severe weather events across US.

Event_Type

CASUALTIES

TOTAL_FATALITIES

TOTAL_INJURIES

1

TORNADO

96979.00

5633.00

91346.00

2

THUNDERSTORM WIND

10054.00

701.00

9353.00

3

EXCESSIVE HEAT

8428.00

1903.00

6525.00

4

FLOOD

7259.00

470.00

6789.00

5

LIGHTNING

6046.00

816.00

5230.00

6

HEAT

3037.00

937.00

2100.00

7

FLASH FLOOD

2755.00

978.00

1777.00

8

ICE STORM

2064.00

89.00

1975.00

9

WILDFIRE

1543.00

87.00

1456.00

10

WINTER STORM

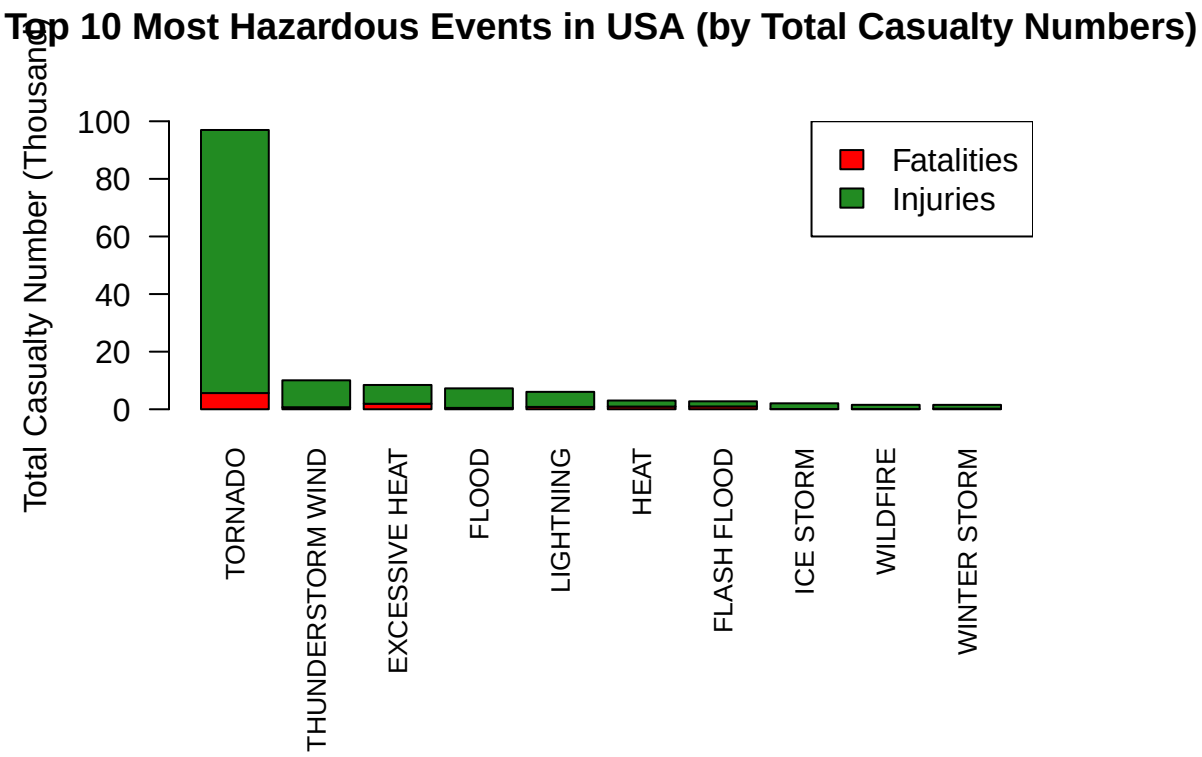
1527.00

206.00

1321.00

We genereate a barplot to have an overview of the result.

```
par(mar=c(10,5,5,5))
barplot(my_cas_barplot_data,col=c("red","forestgreen"),names.arg = as.character(sum_cas_stormdata[1:10,]
title(main = "Top 10 Most Hazardous Events in USA (by Total Casualty Numbers)",ylab = "Total Casualty Nu
legend("topright",c("Fatalities","Injuries"),fill=c("red","forestgreen"))
```



The tonados caused the most casualties in US according to the data. Which is far more harmful than the second and third ones. The reason might be tonado is usually highly unpredictable. However we shouldn't overlook other severe weather events some of them may not be so unpredictable still precautions should be taken in advance.

(**Anser to Q2**)
Below is the top 10 most distructive severe weather events across US.

```
report_eco_stormdata<-sum_eco_stormdata[1:10,]
xreport_eco<-xtable(select(report_eco_stormdata,Event_Type,PROPERTY_LOSS,CROP_LOSS_M,TOTAL_LOSS),captio
print.xtable(xreport_eco,type = "html")
```

Top 10 most distructive severe weather events across US

Event_Type

PROPERTY_LOSS

CROP_LOSS_M

TOTAL_LOSS

1

TORNADO

51.64

0.41

51.64

2

FLOOD

22.16

5.66

22.16

3

FLASH FLOOD

15.14

1.42

15.14

4

HAIL

13.93

3.03

13.94

5

HURRICANE/TYPHOON

9.97

3.84

9.98

6

THUNDERSTORM WIND

9.70

1.16

9.71

7

WILDFIRE

5.23

0.40

5.23

8

HIGH WIND

3.97

0.64

3.97

9

ICE STORM

3.94

0.02

3.94

10

TROPICAL STORM

2.55

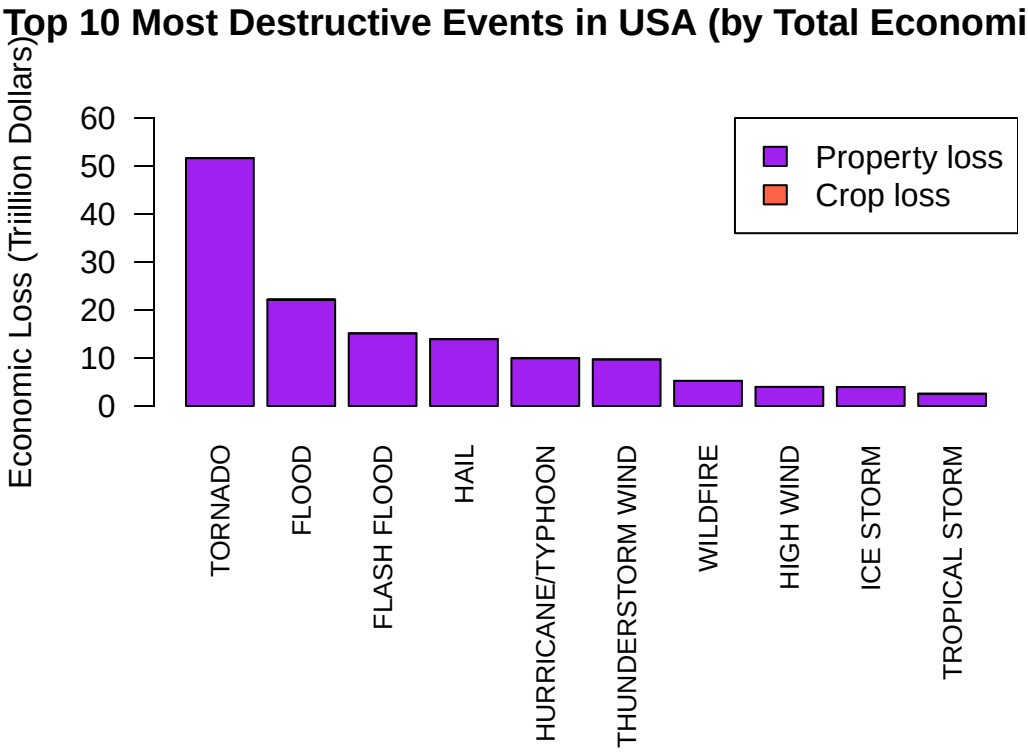
0.68

2.55

(CROP_LOSS is in million dollars others are in trillion dollars)

We genereate a barplot to have an overview of the result.

```
par(mar=c(10,5,5,5))
barplot(my_eco_barplot_data,col=c("purple","tomato"),names.arg = as.character(sum_eco_stormdata[1:10,]$
title(main = "Top 10 Most Destructive Events in USA (by Total Economic Loss )",ylab = "Economic Loss (T
legend("topright",c("Property loss","Crop loss"),fill=c("purple","tomato"))
```



The tonados caused the most economic loss across US. Almost twice as much as the second event, flood. The property damage constitutes the major part of economic loss.