Analysing U.S. NOAA Storm Data

Harish Kumar Rongala

January 16, 2017

1. Synopsis

This data analysis trys to 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?

2. Data Description

Data is provided by U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and property damage.

Data set used in this analysis was downloaded from here Documentation for this data set is available here

3. Data Processing

In this section, we download the data set which is comma seperated value (csv) text file, encrypted with bz2 algorithm. We can read this data using **read.csv** method in R.

```
url<-"https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2";
download.file(url,"storm_data.csv.bz2");
raw_data<-read.csv("storm_data.csv.bz2");
```

Let's take a glance at the loaded data set named **raw_data**.

```
print(dim(raw_data));
```

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

```
print(names(raw_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"
```

4. Data Transformation

There are 37 columns in the data set. To answer our questions we may not need all these columns. For each question, we have to consider different set of variables.

4.1. Transformation for first question

To answer the first question we have to consider "FATALITIES" & "INJURIES" to weight the harmful effects of each "EVTYPE"(Tornado, Flood etc) towards population health.

```
## Create new data frame filtering unwanted variables
first<-data.frame(raw_data$EVTYPE,raw_data$FATALITIES,raw_data$INJURIES);
names(first)<-c("EventType","Fatalities","Injuries");
## Aggregating total fatalities by Event type
fatalities<-aggregate(first$Fatalities,by=list(first$Event),FUN=sum);
names(fatalities)<-c("Event","Total_Fatalities");
## Order the fatalities in descending order
fatalities<-fatalities[order(fatalities$Total_Fatalities,decreasing=T),];
## Aggregating total injuries by Event type
injuries<-aggregate(first$Injuries,by=list(first$Event),FUN=sum);
names(injuries)<-c("Event","Total_Injuries");
## Order the injuries in descending order
injuries<-injuries[order(injuries$Total_Injuries,decreasing=T),];
```

4.2. Transformation for second question

To answer the second question, we consider "EVTYPE", "PROPDMG","PROPDMGEXP","CROPDMG" and "CROPDMGEXP", because they represent the **Economic impact**.

```
second<-data.frame(raw_data$EVTYPE,raw_data$PROPDMG,raw_data$PROPDMGEXP,raw_data$CROPDMG,raw_data$CROPD
names(second)<-c("Event","PropertyDmg","PropertyDmgExp","CropDmg","CropDmgExp");
```

"PropertyDmgExp" and "CropDmgExp" represents the exponents of "PropertyDmg" and "CropDmg" respectively. Let's look at these exponent values

```
print(levels(second$PropertyDmgExp));
```

```
##  [1] ""  "-" "?" "+" "0" "1" "2" "3" "4" "5" "6" "7" "8" "B" "h" "H" "K"
## [18] "m" "M"
```

```
print(levels(second$CropDmgExp));
```

```
## [1] ""  "?" "0" "2" "B" "k" "K" "m" "M"
```

According to the documentation, provided by U.S.N.O.A.A. Each non-numeric level should be interpreted as below

Symbol	Interpretation
"" "?" "-" "+"	0
"h" "H"	2
"k" "K"	3
"m" "M"	6
"b" "B"	9

Let's convert these non-numerical levels. First update the 'PropertyDmgExp' column.

```
## Assign zero for miscellaneous values
levels(second$PropertyDmgExp)[levels(second$PropertyDmgExp)=="" | levels(second$PropertyDmgExp)=="-" | l
## Assign 9 for Billion
levels(second$PropertyDmgExp)[levels(second$PropertyDmgExp)=="B"]<-"9";
```

```
## Assign 6 for Million
levels(second$PropertyDmgExp)[levels(second$PropertyDmgExp)=="m" | levels(second$PropertyDmgExp)=="M"]<
## Assign 3 for thousand's
levels(second$PropertyDmgExp)[levels(second$PropertyDmgExp)=="K"]<-"3";
## Assign 2 for hundred's
levels(second$PropertyDmgExp)[levels(second$PropertyDmgExp)=="h" | levels(second$PropertyDmgExp)=="H"]<
```

Now, update 'CropDmgExp' column.

```
## Assign zero for miscellaneous values
levels(second$CropDmgExp)[levels(second$CropDmgExp)=="?" | levels(second$CropDmgExp)==""]<-"0";
## Assign 3 for thousand's
levels(second$CropDmgExp)[levels(second$CropDmgExp)=="k" | levels(second$CropDmgExp)=="K"]<-"3";
## Assign 6 for Million
levels(second$CropDmgExp)[levels(second$CropDmgExp)=="m" | levels(second$CropDmgExp)=="M"]<-"6";
## Assign 9 for Billion
levels(second$CropDmgExp)[levels(second$CropDmgExp)=="B"]<-"9";
```

Now, in order to get the actual damage of 'Property' and 'Crop', we exponentiate the 'PropertyDmg', 'CropDmg' with 'PropertyDmgExp', 'CropDmgExp' respectively.

```
## Handling Property data
prop<-data.frame(second$Event,second$PropertyDmg,second$PropertyDmgExp);
names(prop)<-c("Event","PropertyDmg","PropertyDmgExp");
prop$total<-prop$PropertyDmg*10**as.numeric(as.character(prop$PropertyDmgExp));
prop<-prop[order(prop$total,decreasing=T),];

## Handling Crop data
crop<-data.frame(second$Event,second$CropDmg,second$CropDmgExp);
names(crop)<-c("Event","CropDmg","CropDmgExp");
crop$total<-crop$CropDmg*10**as.numeric(as.character(crop$CropDmgExp));
crop<-crop[order(crop$total,decreasing=T),];

## Aggregate crop and property data
crop_agg<-aggregate(crop$total,by=list(crop$Event),FUN=sum);
prop_agg<-aggregate(prop$total,by=list(prop$Event),FUN=sum);
names(crop_agg)<-c("Event","Total_Crop");
names(prop_agg)<-c("Event","Total_Prop");

## Re-order in decreasing order of total damage
crop_agg<-crop_agg[order(crop_agg$Total_Crop,decreasing = T),];
prop_agg<-prop_agg[order(prop_agg$Total_Prop,decreasing = T),];

## Merge crop and property data by "Event"
comm<-merge(prop_agg,crop_agg,by="Event");
```

Data set is transformed as required to address our questions.

5. Results

5.1. Population Health

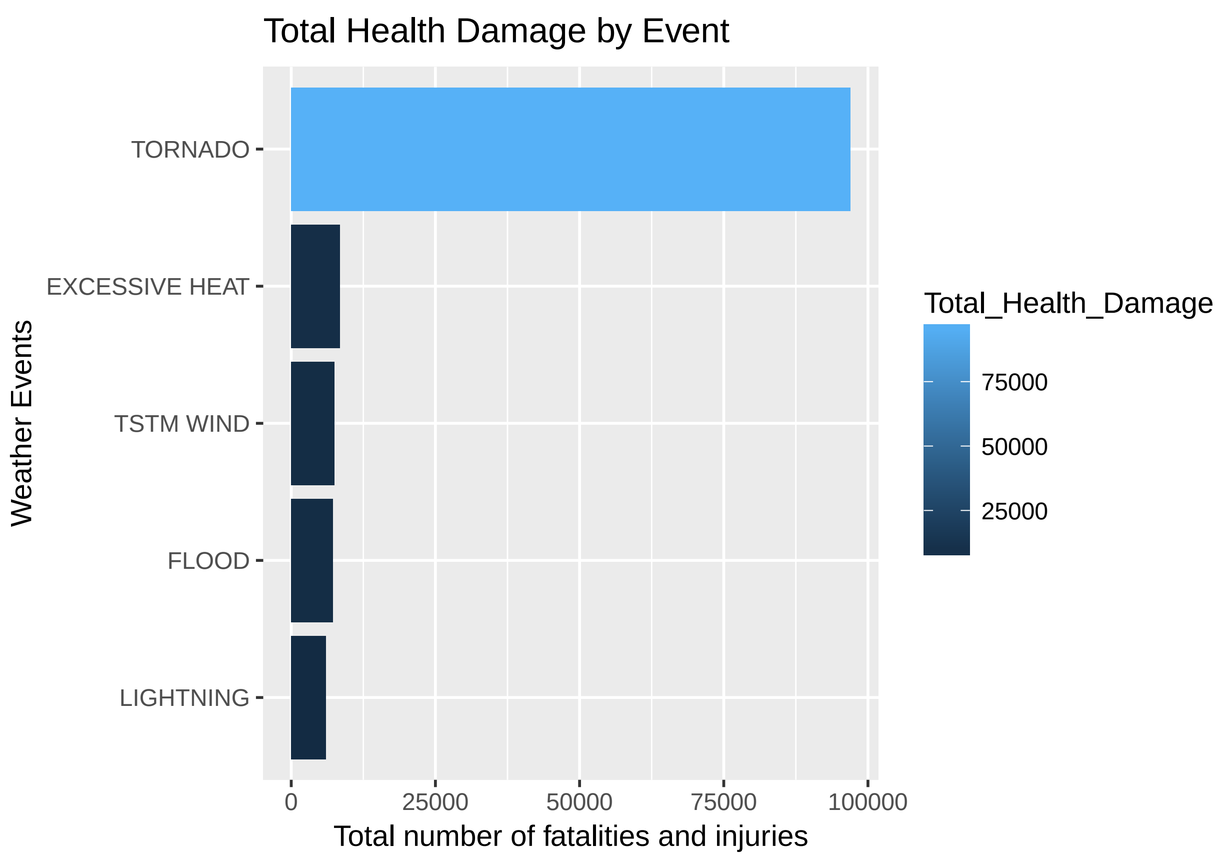
Let's look at the **FATALITIES** and **INJURIES** individually.

```r
head(cbind(fatalities,injuries));
```

```
##              Event Total_Fatalities          Event Total_Injuries
## 834        TORNADO             5633        TORNADO          91346
## 130 EXCESSIVE HEAT             1903      TSTM WIND           6957
## 153    FLASH FLOOD              978          FLOOD           6789
## 275           HEAT              937 EXCESSIVE HEAT           6525
## 464      LIGHTNING              816      LIGHTNING           5230
## 856      TSTM WIND              504           HEAT           2100
```

Above table shows that **TORNADO** is the top most event which has both highest number of fatalities and injuries. However, fatalities and injuries together contribute to **population health**. So, let's merge and re-order them.

```r
## Load 'ggplot' for plotting
library(ggplot2);
## Merge fatalities and injuries
first<-merge(injuries,fatalities);
## Add up total fatalities and injuries in to 'Total_Health_Damage'
first$Total_Health_Damage<-first$Total_Fatalities+first$Total_Injuries;
## Order the data set in the descending order of 'Total_Health_Damage'
first<-first[order(first$Total_Health_Damage,decreasing=T),];
## Plot top 5 harmful events and their impact
first_plot<-ggplot(head(first,5),aes(x=reorder(Event,Total_Health_Damage),y=Total_Health_Damage,fill=To
print(first_plot);
```



Not surprisingly, **Tornado** is the major weather event that has huge impact on population health, resulting in 5633 fatalities and 91.346K injuries.

5.2. Economical Damage

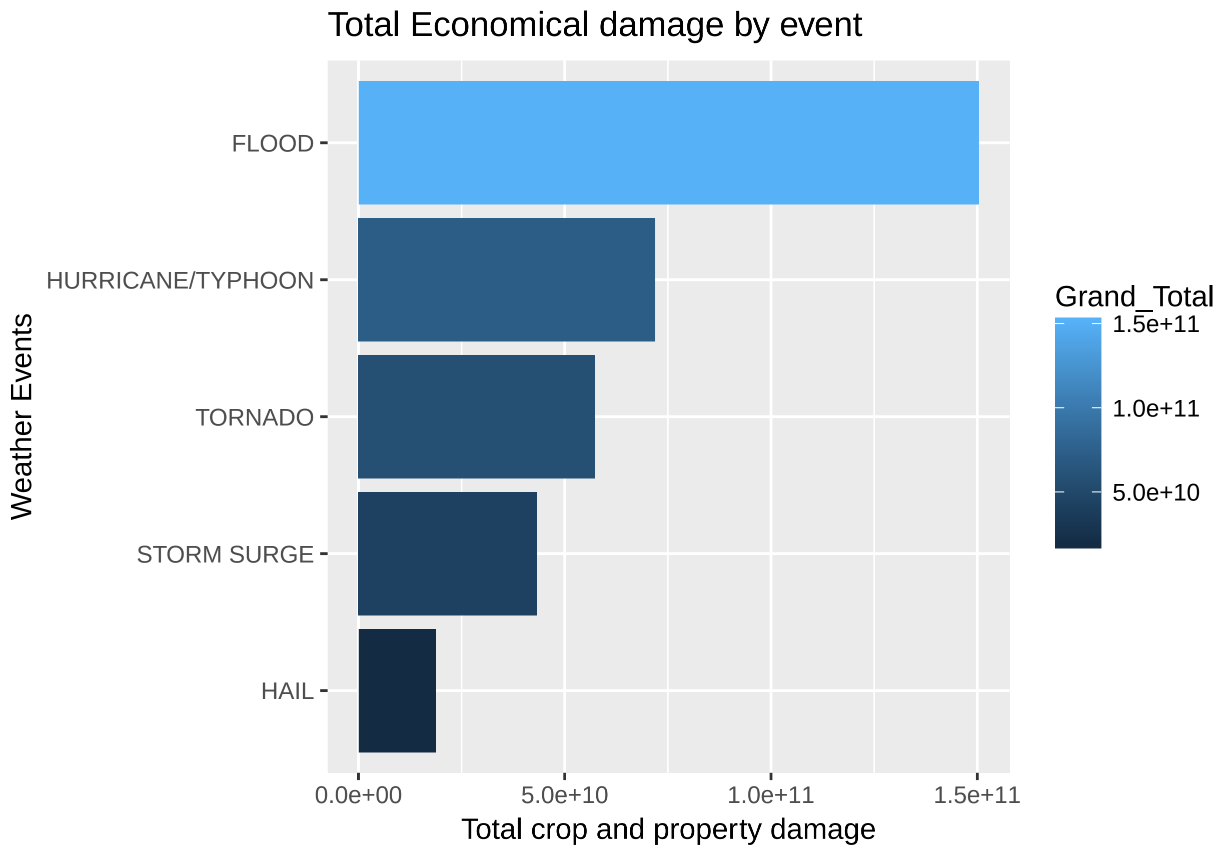
Now let's individually look at the crop and property damage.

```
## Individually look at the crop and property damages
head(cbind(crop_agg,prop_agg));
```

```
##           Event  Total_Crop             Event    Total_Prop
## 95      DROUGHT 13972566000             FLOOD 144657709807
## 170       FLOOD  5661968450 HURRICANE/TYPHOON  69305840000
## 590 RIVER FLOOD  5029459000           TORNADO  56947380677
## 427   ICE STORM  5022113500       STORM SURGE  43323536000
## 244        HAIL  3025954473       FLASH FLOOD  16822673979
## 402   HURRICANE  2741910000              HAIL  15735267513
```

From the above table, **Drough** is the major event when we consider Crop damage. **Flood** is the major event when we consider property damage. However, both crop and property damage contribute towards Economical loss.

```
## Add up crop damage and property damage
comm$Grand_Total<-comm$Total_Prop+comm$Total_Crop;
comm<-comm[order(comm$Grand_Total,decreasing = T),];
library(ggplot2);
second_plot<-ggplot(head(comm,5),aes(x=reorder(Event,Grand_Total),y=Grand_Total,fill=Grand_Total))+geom_
print(second_plot);
```



Flood seems to be the weather event that caused major Economical loss. However, if we consider crop loss alone towards Economical loss, then the weather event responsible will be **drought**.