Reproducible research :Project 2

Mrugank Akarte

24 January 2016

EXPLORING STORM DATA

Synopsis

In this data analysis "Storm Data" provided by 'U.S. National Oceanic and Atmospheric Administration' is used to answer following questions.

1.Which type of events are most harmful with respect to population health?

2.Which type of events have greatest economic Consequences?

Initially this big dataset is made smaller by subsetting database for ease of calculations and further analysis. The various variables considered are 'EVTYPE', 'INJURIES', 'FATALITIES', 'PROPDMG', 'PROPDMGEXP', 'CROPDMG', 'CROPDMGEXP'. Various plots describing the relation between events and injuries, events and fatalities, events and property damaged are plotted. At the end it is concluded that '**TORNADOES**' are most harmful because, maximum injuries and fatalities are caused due them. Also, Tornadoes have greatest economic consequences due to maximum property destruction.

The dataset was downloaded from the following website.

[https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2]

Loading and processing Data

```
df<-read.csv(bzfile("repdata-data-StormData.csv.bz2"))
head(df)
```

```
##   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
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    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
## 4 TORNADO         0                                               0
## 5 TORNADO         0                                               0
## 6 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
## 4         NA         0                       0.0   100 2   0          0
```

```
## 5         NA         0                      0.0   150 2   0          0
## 6         NA         0                      1.5   177 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
## 4        2     2.5          K       0
## 5        2     2.5          K       0
## 6        6     2.5          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
## 4     3458      8626          0          0              4
## 5     3412      8642          0          0              5
## 6     3450      8748          0          0              6
```

```
subdf<-df[,c("EVTYPE","INJURIES","FATALITIES","PROPDMG","PROPDMGEXP","CROPDMG","CROPDMGEXP")]
head(subdf)
```

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

Fisrt analyzing data for variables 'injuries' and 'fatalities'.

```
injuries_df<-subdf[,c("EVTYPE","INJURIES","FATALITIES")]
injuries_df<-injuries_df[which(injuries_df$INJURIES>0),]
```

Checking for injuries and displaying the data in descending order i.e in order of events that have caused maximum injuries first.

```
injuries_df1<-aggregate(INJURIES ~ EVTYPE, injuries_df, sum)
injuries_df1<-injuries_df1[order(injuries_df1$INJURIES, decreasing = T),]
head(injuries_df1)
```

```
##             EVTYPE INJURIES
## 129        TORNADO    91346
## 135      TSTM WIND     6957
## 30           FLOOD     6789
## 20  EXCESSIVE HEAT     6525
## 85       LIGHTNING     5230
## 47            HEAT     2100
```

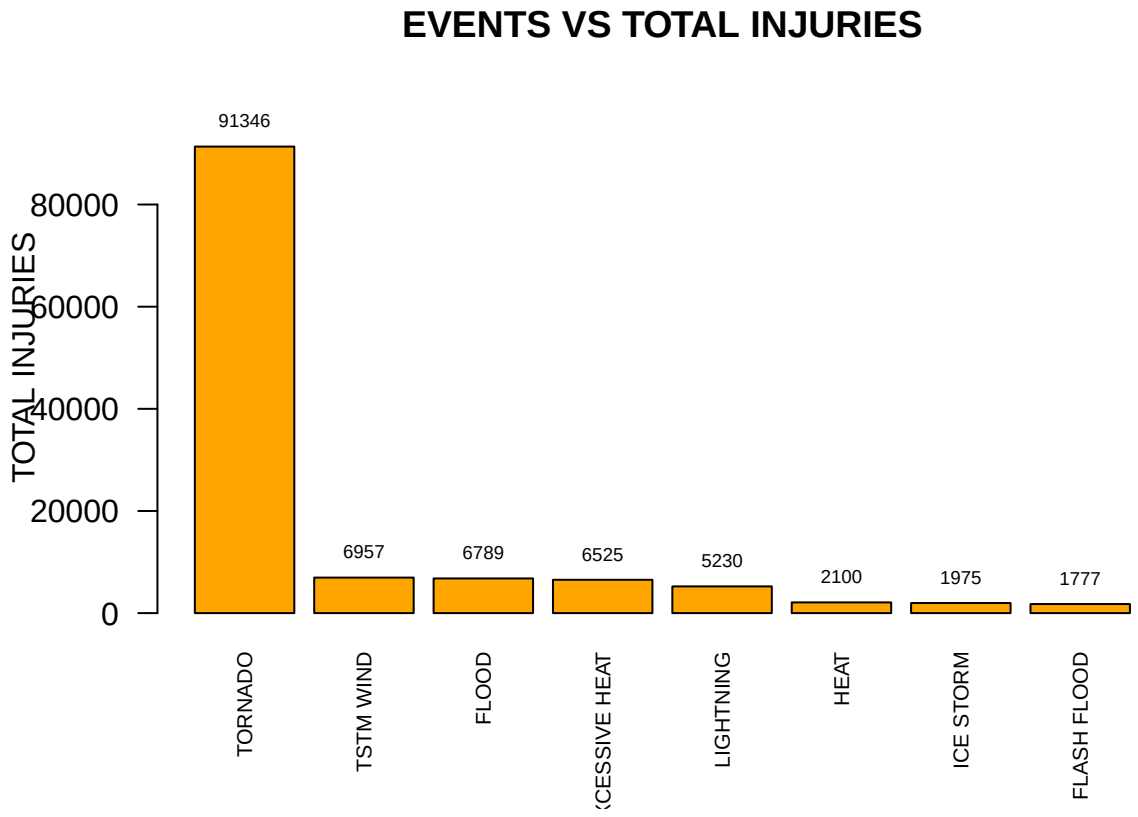
Similarly calculating for Fatalities.

```
injuries_df2<-aggregate(FATALITIES ~ EVTYPE, injuries_df, sum)
injuries_df2<-injuries_df2[order(injuries_df2$FATALITIES, decreasing = T),]
head(injuries_df2)
```

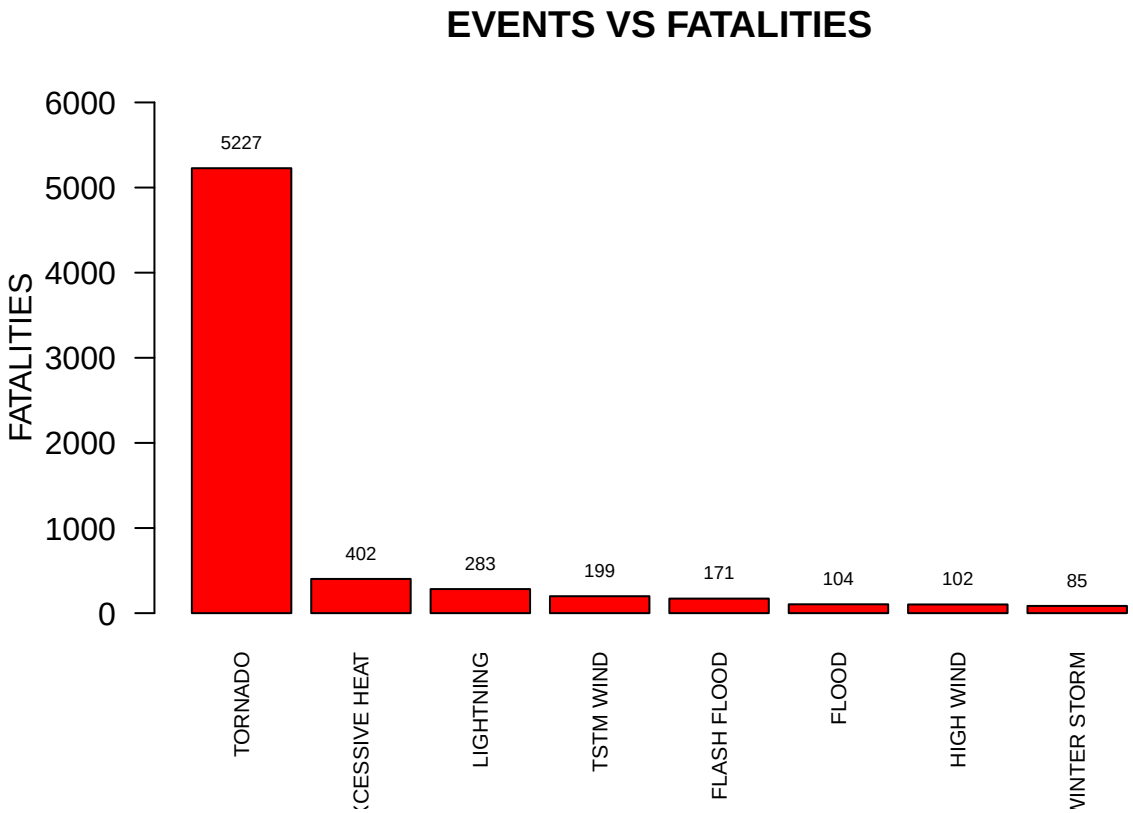
```
##             EVTYPE FATALITIES
## 129        TORNADO       5227
## 20  EXCESSIVE HEAT        402
## 85       LIGHTNING        283
## 135      TSTM WIND        199
## 28     FLASH FLOOD        171
## 30           FLOOD        104
```

Plotting the graphs Events vs Injuries and Events vs Fatalities.

```
plot1<-barplot(injuries_df1$INJURIES[1:8],ylim = c(0,99999),names.arg = injuries_df1$EVTYPE[1:8], cex.na
text(x=plot1, y=injuries_df1$INJURIES[1:8], label= injuries_df1$INJURIES[1:8],cex= 0.6, pos = 3)
```




```
plot2<-barplot(injuries_df2$FATALITIES[1:8],ylim = c(0,6000),names.arg = injuries_df2$EVTYPE[1:8], cex.n
text(x=plot2, y=injuries_df2$FATALITIES[1:8], label= injuries_df2$FATALITIES[1:8],cex= 0.6, pos = 3)
```



From above graphs we can see that **TORNADOES** are most harmful causing maximum injuries and fatalities.

Now analyzing data for property and crop damage.

Before proceeding we need to take care of multipliers used in data in the form of 'k', 'm', 'B'... Thus with the help of function we can calculate total cost of damage in following way.

```
damage_df<-subdf[,c("EVTYPE","PROPDMG","PROPDMGEXP","CROPDMG","CROPDMGEXP")]

func1<-function(value,unit)
{

      if(unit=="h"|| unit=="H"||unit=="2")
      {x<-100}
      else if(unit=="K"||unit=="3")
      {x<-1000}
      else if(unit=="m"||unit=="M"||unit=="6")
      {x<-1000000}
      else if(unit=="B"||unit=="9")
      {x<-1000000000}
      else if(unit=="0")
      {x<-1}
      else if(unit=="1")
      {x<-10}
      else if(unit=="4")
      {x<-10000}
```

```
      else if(unit=="5")
      {x<-100000}
      else if(unit=="7")
      {x<-10000000}
      else if(unit=="8")
      {x<-100000000}
      else
            x<-1

      return(value*x)
}

damage_df$propertydamage<-func1(damage_df$PROPDMG,damage_df$PROPDMGEXP)
damage_df$cropdamage<-func1(damage_df$CROPDMG,damage_df$CROPDMGEXP)
```

Calculating total damage caused by various events.

```
n_pdmg<-aggregate(propertydamage ~ EVTYPE, damage_df, sum)
n_cdmg<-aggregate(cropdamage ~ EVTYPE, damage_df, sum)
```

Organizing the above data in descending order so that we can get clear information about which events have caused maximum property and crop destruction.

```
n_pdmg<-n_pdmg[order(n_pdmg$propertydamage, decreasing = T),]
n_cdmg<-n_cdmg[order(n_cdmg$cropdamage, decreasing = T),]
head(n_pdmg)
```

```
##                EVTYPE propertydamage
## 834           TORNADO     3212258160
## 153       FLASH FLOOD     1420124590
## 856         TSTM WIND     1335965610
## 170             FLOOD      899938480
## 760 THUNDERSTORM WIND      876844170
## 244              HAIL      688693380
```

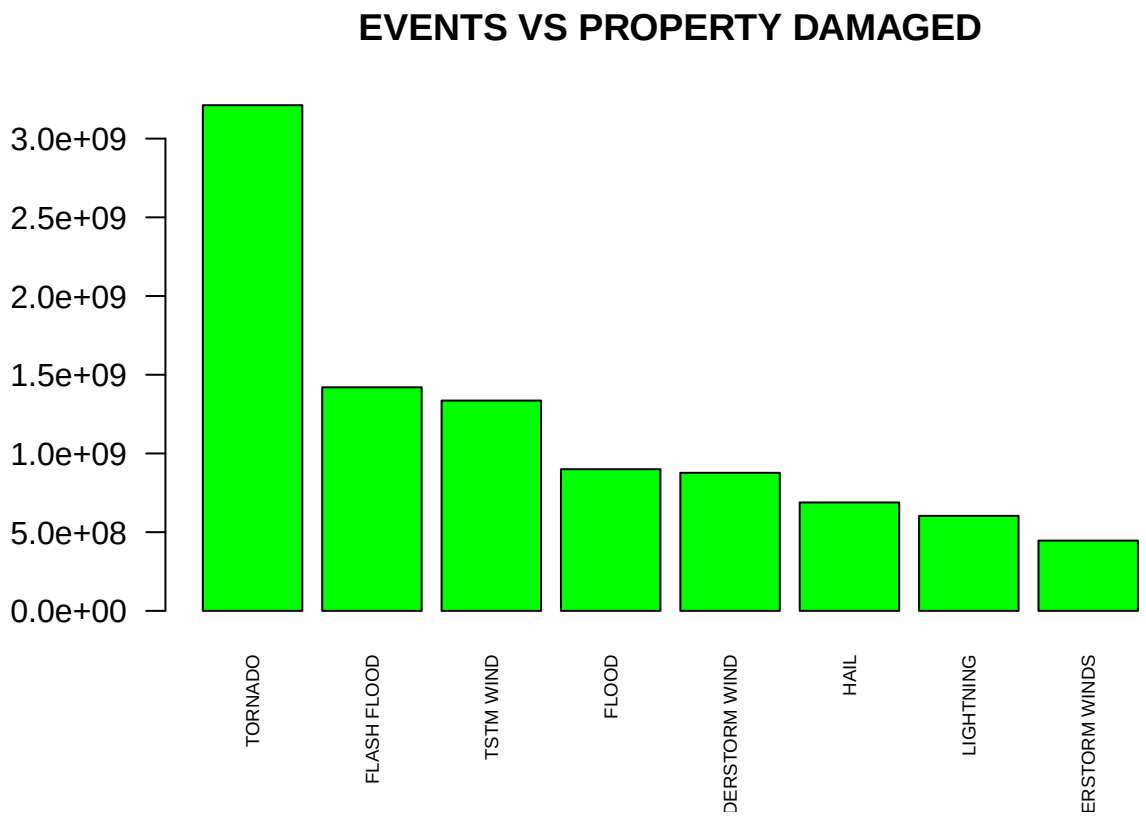
```
head(n_cdmg)
```

```
##                EVTYPE cropdamage
## 244              HAIL  579596.28
## 153       FLASH FLOOD  179200.46
## 170             FLOOD  168037.88
## 856         TSTM WIND  109202.60
## 834           TORNADO  100018.52
## 760 THUNDERSTORM WIND   66791.45
```

Looking at above data we can say that *TORNADOES* are the major cause for damaging property and *HAIL STORM* for damaging crops.

Plotting a barplot of Events vs Property damaged.

```
barplot(n_pdmg$propertydamage[1:8],names.arg = n_pdmg$EVTYPE[1:8], cex.names = 0.6, las=2, col="green",m
```



From the above plot we see that *TORNADOES* are the main reason of maximum property damage.

RESULTS

Thus from above data analysis we can say that **TORNADOES** are most harmful with respect to *human population* since tornadoes have caused maximum injuries and fatalities. They are also major cause for *destroying property* while **HAIL STORM** is the major reason for *crop destruction*.