Ananlysis of Storm dataset with respect to health and economic damage

Summary

The data comes from the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database and contains data from 1950 to November 2011. We are interested to study and compare the damage caused by storms and other weather events in the United States in the given period. The analysis involves reading the data in the software and applying necessary transformations(discussed in the data processng section) on the data and comparing the damage by means of graphical aid.

Data Processing

Steps used to format and group data.

1. The data is loaded into R using the read.csv function.
2. The columns named "EVTYPE","FATALITIES","INJURIES","PROPDMG","CROPDMG","PROPDMGEXP","CROPDMGEXP" are extracted from the data as they are the data of interest.
3. The names of the columns are changed to lower case for ease of writing code.

```
# reading and formatting data

data=read.csv("C:\\Users\\HP\\R\\repdata_data_StormData.csv\\repdata_data_StormData.csv")
names(data)=tolower(names(data))
data=data[,c("evtype","fatalities","injuries","propdmg","cropdmg")]
head(data)
```

```
##    evtype fatalities injuries propdmg cropdmg
## 1 TORNADO          0       15    25.0       0
## 2 TORNADO          0        0     2.5       0
## 3 TORNADO          0        2    25.0       0
## 4 TORNADO          0        2     2.5       0
## 5 TORNADO          0        2     2.5       0
## 6 TORNADO          0        6     2.5       0
```

4. We group the data by "EVTYPE" and calculate the following:
    - sumh=stores the sum of total fatalities and injuries for each "EVTYPE".
    - sump=stores the sum of total property damage for each "EVTYPE".
    - sumc=stores the sum of total crop damage for each "EVTYPE".

```
library(dplyr)
data_dmg=data %>%
  group_by(evtype) %>%
  summarise(sumh=sum(fatalities)+sum(injuries),sump=sum(propdmg),sumc=sum(cropdmg))

data_dmg
```

```
## # A tibble: 985 x 4
##    evtype                   sumh   sump  sumc
##    <fctr>                  <dbl>  <dbl> <dbl>
##  1 "   HIGH SURF ADVISORY"     0 200         0
##  2 " COASTAL FLOOD"            0   0         0
##  3 " FLASH FLOOD"              0  50.0       0
##  4 " LIGHTNING"                0   0         0
##  5 " TSTM WIND"                0 108         0
##  6 " TSTM WIND (G45)"          0   8.00      0
##  7 " WATERSPOUT"               0   0         0
##  8 " WIND"                     0   0         0
##  9 ?                           0   5.00      0
## 10 ABNORMAL WARMTH             0   0         0
## # ... with 975 more rows
```

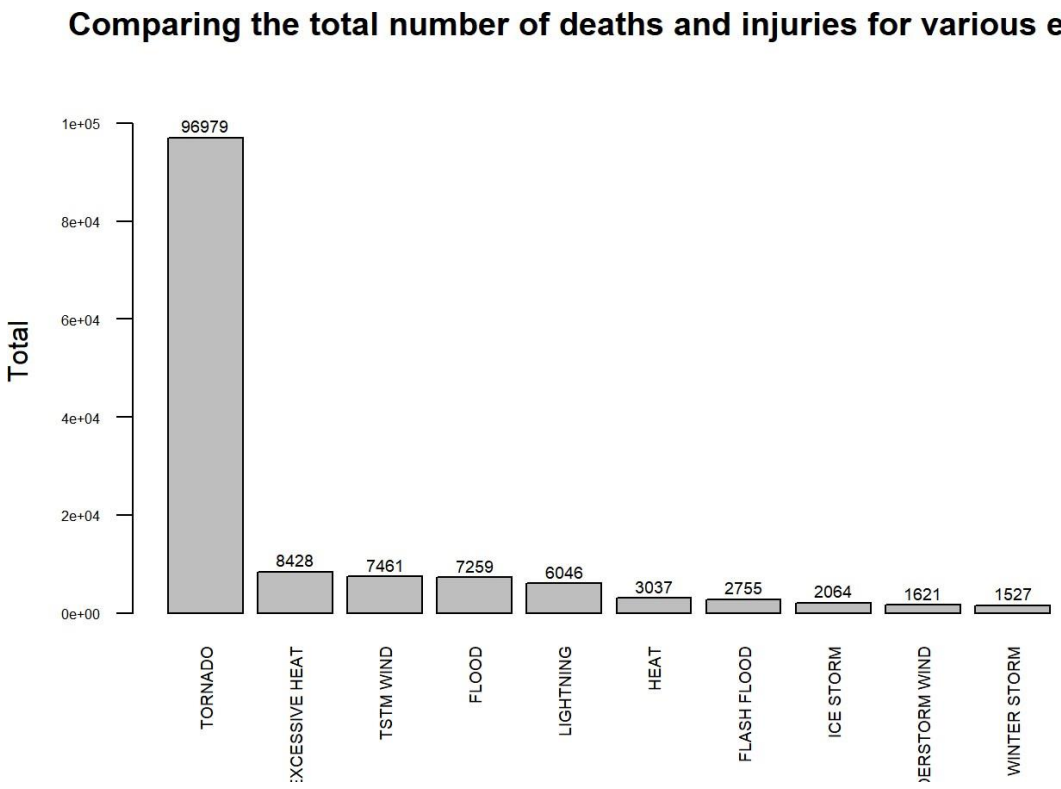
Results

Due to presence of large number of missing data and untidy data, we order the data in decreasing order of the number of fatalities+injuries and plot the first 10 values in a barplot.

```
data_health=data_dmg[order(data_dmg$sumh,decreasing = TRUE),]
data_health=data_health[1:10,]
data_health[,c(1,2)]
```

```
## # A tibble: 10 x 2
##    evtype             sumh
##    <fctr>            <dbl>
##  1 TORNADO           96979
##  2 EXCESSIVE HEAT     8428
##  3 TSTM WIND          7461
##  4 FLOOD              7259
##  5 LIGHTNING          6046
##  6 HEAT               3037
##  7 FLASH FLOOD        2755
##  8 ICE STORM          2064
##  9 THUNDERSTORM WIND  1621
## 10 WINTER STORM       1527
```

```
hp=with(data_health,barplot(sumh,names.arg=evtype,ylab="Total",
                  ylim=c(0,1.1*max(data_health$sumh)),
                  main="Comparing the total number of deaths and injuries for various events",
                  cex.names=0.6,las=2,cex.axis=0.5))
text(hp,data_health$sumh+2500,labels=data_health$sumh,cex=0.6,col="black")
```



Comments

Based on the graph, we can say that Tornado is most hazardous with respect to population health with other events contributing very little to the population health hazard.

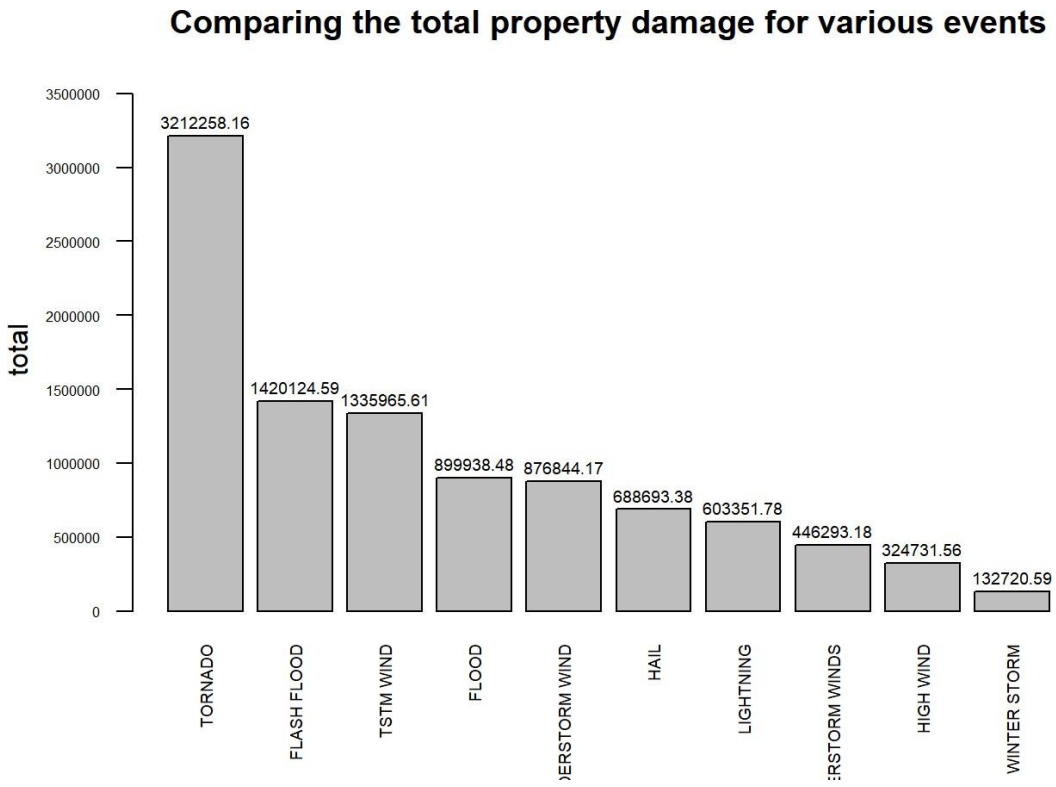
==========

Due to presence of large number of missing data and untidy data, we order the data in decreasing order of the number of property damage and plot the first 10 values in a barplot.

```
data_prop=data_dmg[order(data_dmg$sump,decreasing=TRUE),]
data_prop=data_prop[1:10,]
data_prop[,c(1,3)]
```

```
## # A tibble: 10 x 2
##    evtype                sump
##    <fctr>               <dbl>
##  1 TORNADO            3212258
##  2 FLASH FLOOD        1420125
##  3 TSTM WIND          1335966
##  4 FLOOD               899938
##  5 THUNDERSTORM WIND   876844
##  6 HAIL                688693
##  7 LIGHTNING           603352
##  8 THUNDERSTORM WINDS  446293
##  9 HIGH WIND           324732
## 10 WINTER STORM        132721
```

```
pp=with(data_prop,barplot(sump,names.arg=evtype,ylab="total",
                          ylim=c(0,1.1*max(data_prop$sump)),
                         main="Comparing the total property damage for various events",
                          cex.names=0.6,las=2,cex.axis=0.5))
text(pp,data_prop$sump+95000,labels=data_prop$sump,cex=0.6,col="black")
```



Comments

For the total property damage by an event, Tornadoes have the highest damage cost. Floods and TSTM winds having almost similar damage cost.

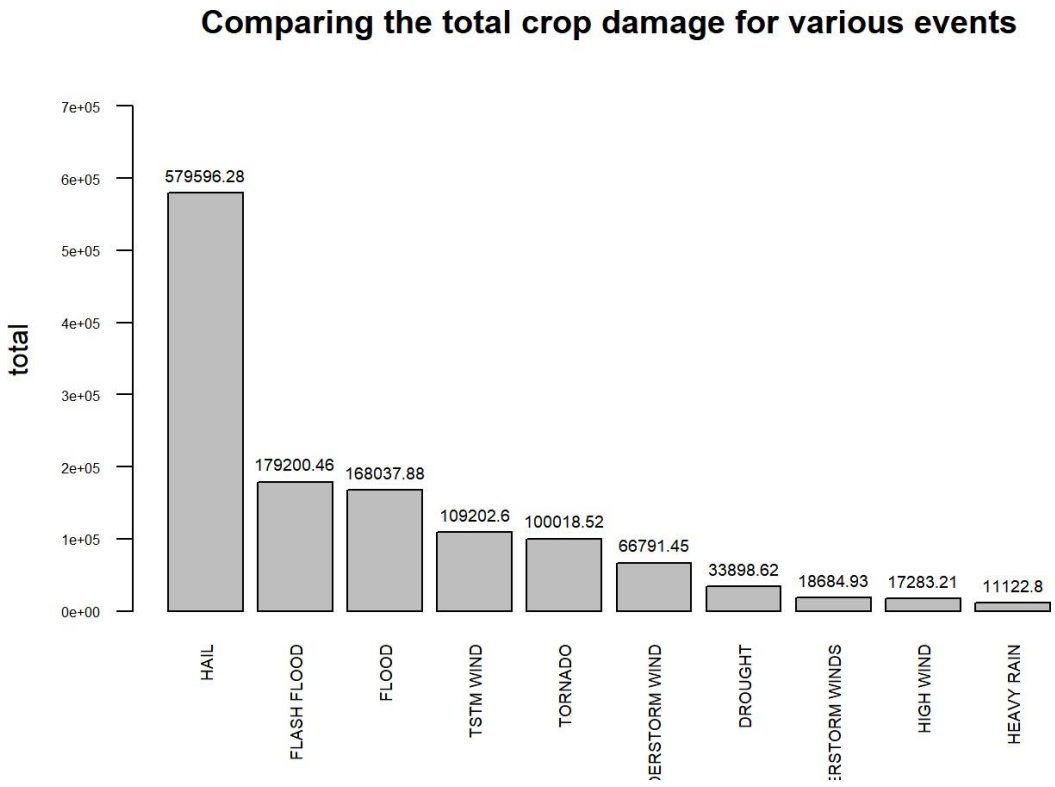
==========

Due to presence of large number of missing data and untidy data, we order the data in decreasing order of the number of crop damage and plot the first 10 values in a barplot.

```
data_crop=data_dmg[order(data_dmg$sumc,decreasing=TRUE),]
data_crop=data_crop[1:10,]
data_crop[,c(1,4)]
```

```
## # A tibble: 10 x 2
##    evtype               sumc
##    <fctr>              <dbl>
##  1 HAIL               579596
##  2 FLASH FLOOD        179200
##  3 FLOOD              168038
##  4 TSTM WIND          109203
##  5 TORNADO            100019
##  6 THUNDERSTORM WIND   66791
##  7 DROUGHT             33899
##  8 THUNDERSTORM WINDS  18685
##  9 HIGH WIND           17283
## 10 HEAVY RAIN          11123
```

```
with(data_crop,barplot(sumc,names.arg=evtype,ylab="total",
                       ylim=c(0,1.25*max(data_crop$sumc)),
                     main="Comparing the total crop damage for various events",
                     cex.names=0.6,las=2,cex.axis=0.5))
text(pp,data_crop$sumc+25000,labels=data_crop$sumc,cex=0.6,col="black")
```



Comments

Hail caused the most crop damage over the years. Floods and Flash floods being the second and third most destructive respectviely.

==========

NOTE

- The data had many inconsistencies with respect to the names of the events and require further cleaning.
- The varaibles “propdmgexp” and “cropdmgexp” havent been taken into account while calculating the total damage due to lack of proper definitons.
- The actual interpretation might change when the above factors have been taken into account.