Data analysis to determine weather events with the most harm and economic impact

Summary

The present document explores the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, in order to acknowledge the top harmful and fatal weather events ocurred across the U.S, as well as the ones with the most economic impact.

This analysis project pretends to illustrate the top 6 weather events in with fatal and injury casualties ocurred, as well as the top 6 weather events which had the most economic impact across the country.

Data Processing

The U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database comes in a compressed (.bz2) csv file. The file has to be stored in a dataset:

```
data <- read.csv("repdata_data_StormData.csv.bz2", header = TRUE)
```

The database was subset into an injury/fatality analysis, and an economic impact analysis.

Injury/Fatality weather event based subset:

```
mostHarmful <-with(data, aggregate(INJURIES + FATALITIES ~ EVTYPE, data=data, FUN = "sum"))
```

Economic impact weather event based subset:

```
data2 <- subset(data, (data$CROPDMGEXP == "M" | data$CROPDMGEXP == "K" | data$CROPDMGEXP == "B") | (data
```

Data Analyisis

Once the processed data is ready, the analysis for harmful casualties is described next:

The harmful/fatalities subset must be ordered by a descending order, in terms of casualties, to find the top 5 weather events with the highest number of casualties:

```
mostHarmful <-mostHarmful[order(-mostHarmful[2]),]
newmostHarmful <- mostHarmful[1:5,]
names(newmostHarmful)[2] <- "Casualties"
newmostHarmful
```

```
##             EVTYPE Casualties
## 834        TORNADO      96979
## 130 EXCESSIVE HEAT       8428
## 856      TSTM WIND       7461
## 170          FLOOD       7259
## 464      LIGHTNING       6046
```

The economic impact subset must be processed in a way that the economic impact variable should be a numeric variable. In order to achieve this, the variables *CROPDMGEXP* and *PROPDMGEXP* have a char variable (K for thpusands, M for millons and B for billions), knowing this, the observations from which K appears, the variables *CROPDMG* and *PROPDMG* will be multiplied by 1, the ones where M appears will be multiplied by 1000 and the ones ehere B appears will be multiplied by 1000000:

```
for(i in 1:length(data2$CROPDMGEXP)) {
  ifelse(data2$CROPDMGEXP[i] == "M", data2$CROPDMG[i] <- data2$CROPDMG[i] *  100,
         ifelse(data2$CROPDMGEXP[i] == "K", data2$CROPDMG[i] <- data2$CROPDMG[i] * 1,
                ifelse(data2$CROPDMGEXP[i] == "B", data2$CROPDMG[i] <- data2$CROPDMG[i] * 1000000, data
}


for(i in 1:length(data2$PROPDMGEXP)) {
  ifelse(data2$PROPDMGEXP[i] == "M", data2$PROPDMG[i] <- data2$PROPDMG[i] * 100,
         ifelse(data2$PROPDMGEXP[i] == "K", data2$PROPDMG[i] <- data2$PROPDMG[i] * 1,
                ifelse(data2$PROPDMGEXP[i] == "B", data2$PROPDMG[i] <- data2$PROPDMG[i] * 1000000, data
}
```

Once this processing is ready, the resulting dataset must be ordered by a descending order, in terms of the economic lost, to find out the top 5 weather events which has the most economic and material damage:

```
monetizedDamage <- with(data2, aggregate(CROPDMG + PROPDMG ~ EVTYPE, data=data, FUN = "sum"))
names(monetizedDamage)[2] <- "moneyDamage"
monetizedDamage <- monetizedDamage[order(-monetizedDamage$moneyDamage),]
printableDamage <- head(monetizedDamage,5)
printableDamage
```

```
##          EVTYPE moneyDamage
## 834     TORNADO     3312277
## 153 FLASH FLOOD     1599325
## 856   TSTM WIND     1445168
## 244        HAIL     1268290
## 170       FLOOD     1067976
```

Results

After the casualties analysis is donde, the results clearly shows that the weather events which, unfortunately, causes the most fatalies and injuries across the U.S is the **TORNADO**, followed by the excessive heat.

```
library(ggplot2)
print(ggplot(data=newmostHarmful, aes(x=factor(EVTYPE, level = EVTYPE), y= Casualties) ) + geom_bar(sta
```

On the other hand, the analysis throws that the event which causes the most economic impact across the U.S is also the **TORNADO**, followed by the Flash Flood.

```
print(ggplot(data=printableDamage, aes(x=factor(EVTYPE, level = EVTYPE), y= moneyDamage) ) + geom_bar(s
```

As a conclusion, the tornado is the most dangerous weather event to impact the United States.

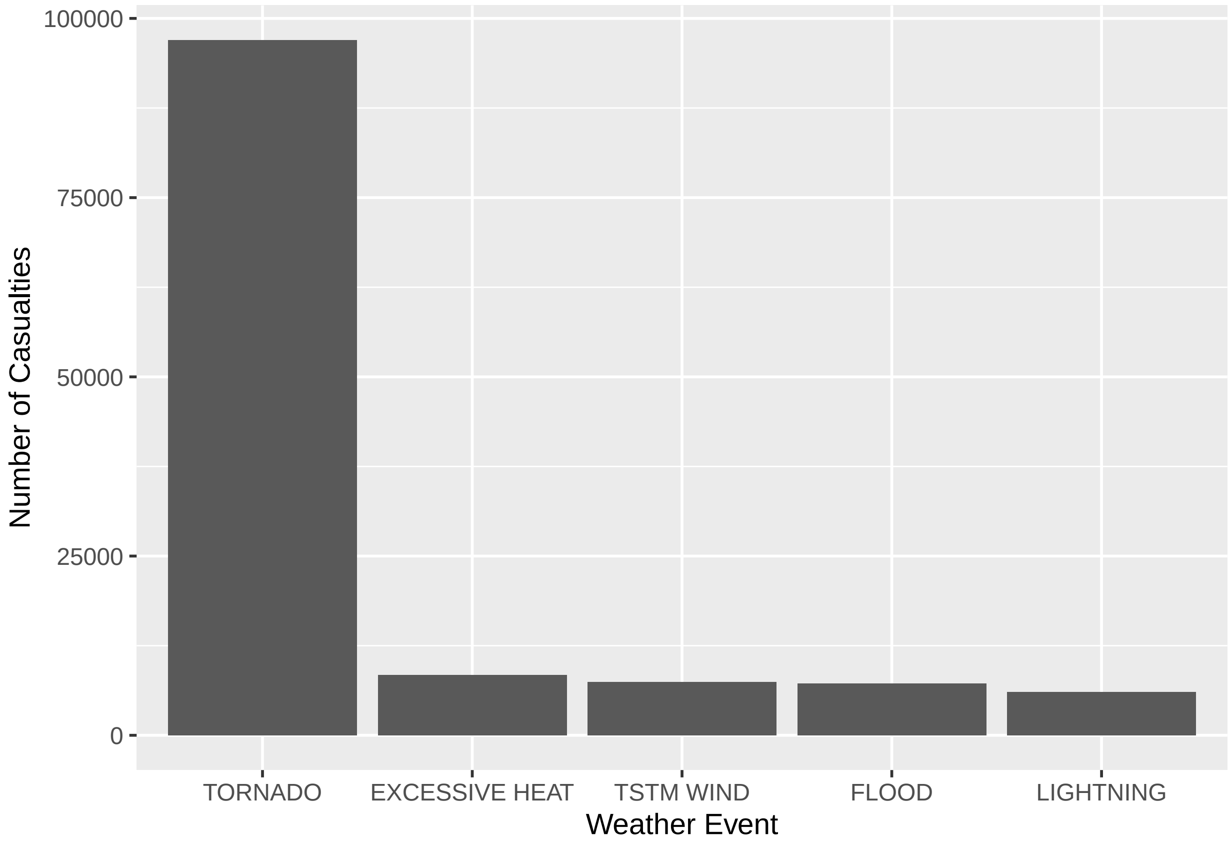

Figure 1: Casualties ocurrences by weather event

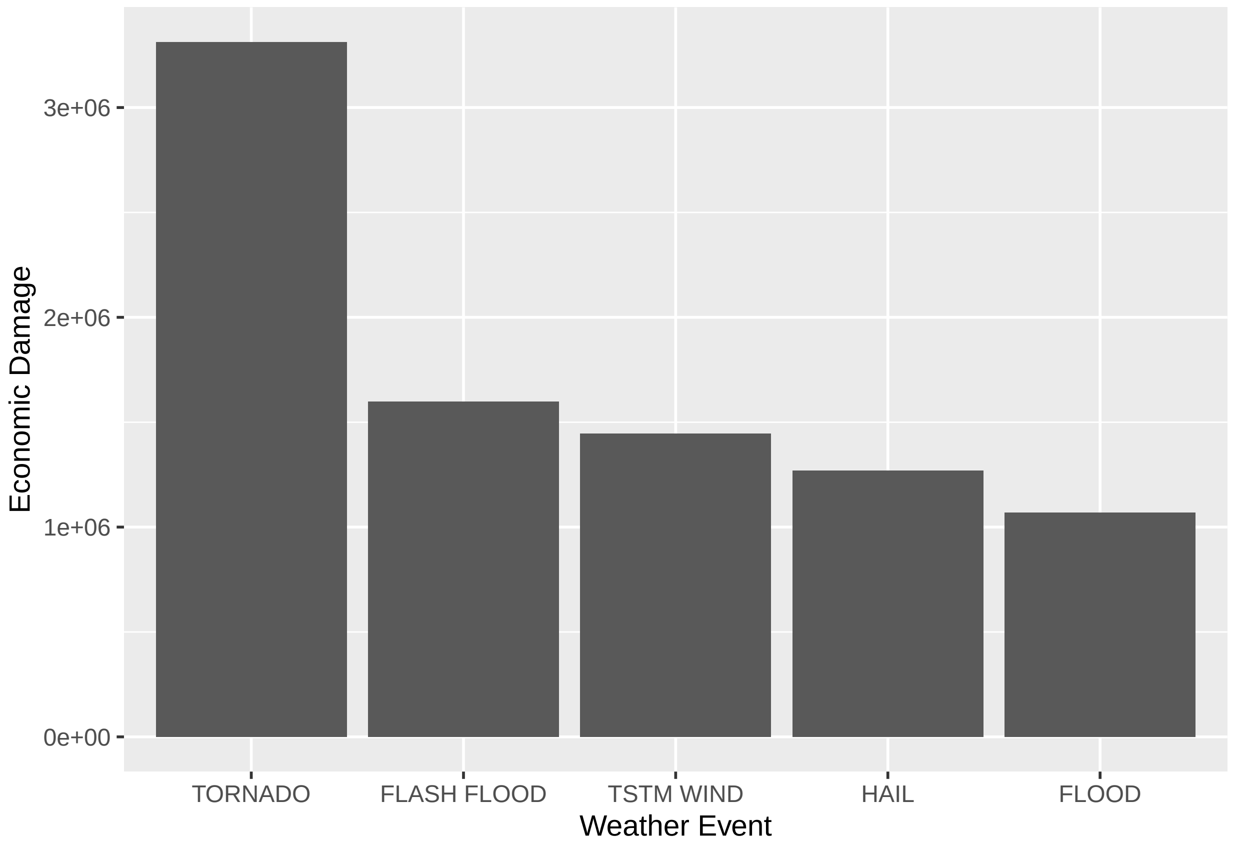

Figure 2: Economic impact by weather event