Damage Analysis on Natural Disasters based on NOAA's Storm Database

Shuming Dou

1/18/2021

Section 1: Synopsis

The objective of this project is to investigate U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database, to find which type of event has the greatest impact on population health, and poses the most severe economic consequences. We begin this analysis with loading data, and then extracting useful columns to form a clean dataset ready for further analysis. Then, we use aggregate() function to find the average fatalities, injuries, property damage, crop damage, and total damages by types of events, and extract the highest 5 types of events in each damage categories. Finally, we use the extracted top-5 data frames to plot barplots to communicate our findings.

Section 2: Data Processing

Before data cleanning, we need to load the raw data set from StormData.csv. After data loading is finished, we take a look at the top 6 rows of the raw data set

```r
## data loading
if (!exists("storm.raw")) {
    storm.raw <- read.csv("./data/StormData.csv")
}
head(storm.raw)
```

```
##   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
```

Since we are concerned with the relationship between types of events (**EVTYPE**) and population health (**FATALIITIES** & **INJURIES**) or economic consequences (**PROPDMG** & **CROPDMG**), we need to extract these 5 colmuns, and removing NA values for the purpose of data cleaning.

```
## extract the only 7 columns that we are interested in:
## EVTYPE, FATALITIES, INJURIES, PROPDMG, CROPDMG
storm.interested <- storm.raw[, c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "CROPDMG")]

## remove any rows that have NA values in any of these 7 variables
storm.clean <- storm.interested[!is.na(storm.interested$EVTYPE) &
                                   !is.na(storm.interested$FATALITIES) &
                                   !is.na(storm.interested$INJURIES) &
                                   !is.na(storm.interested$PROPDMG) &
                                   !is.na(storm.interested$CROPDMG), ]

## convert all lowercase letters to uppercase
storm.clean$EVTYPE <- toupper(storm.clean$EVTYPE)

## take a look at the top 6 rows of the clean dataset
head(storm.clean)
```

```
##    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
```

Section 3: Exploratory Analysis

- **Types of Events vs. Population Health**

We can use aggregated() function to find the average fatalities and injuries by types of events, and store them into two new data frame. Combining these two new data frames grants us to reorder the data frame. Therefore, we can extract the top 5 types of events with highest average fatalities and injuries, separately and together. The extracted top-5 data frames are intended to be plotted in Section 4.

```
## find the average fatalities by types of events
storm.FATALITIES <- aggregate(FATALITIES ~ EVTYPE, storm.clean, mean)

## find the average injuries by types of events
storm.INJURIES <- aggregate(INJURIES ~ EVTYPE, storm.clean, mean)
```

```
## combine two data frame
storm.health <- cbind(storm.FATALITIES, storm.INJURIES$INJURIES)
colnames(storm.health)[3] <- "INJURIES"

## display the top 5 types of events with highest average fatalities and injuries, separately and togeth
storm.FATALITIES.top5 <- head(storm.FATALITIES[order(storm.FATALITIES$FATALITIES, decreasing = TRUE), ]
storm.INJURIES.top5 <- head(storm.INJURIES[order(storm.INJURIES$INJURIES, decreasing = TRUE), ], 5)
storm.health.top5 <- head(storm.health[order(storm.health$FATALITIES, storm.health$INJURIES, decreasing

storm.FATALITIES.top5
```

```
##                           EVTYPE FATALITIES
## 766 TORNADOES, TSTM WIND, HAIL  25.000000
## 62                COLD AND SNOW  14.000000
## 775      TROPICAL STORM GORDON   8.000000
## 519      RECORD/EXCESSIVE HEAT   5.666667
## 127               EXTREME HEAT   4.363636
```

```
storm.INJURIES.top5
```

```
##                      EVTYPE INJURIES
## 775 TROPICAL STORM GORDON     43.0
## 872            WILD FIRES     37.5
## 746         THUNDERSTORMW     27.0
## 327    HIGH WIND AND SEAS     20.0
## 585       SNOW/HIGH WINDS     18.0
```

```
storm.health.top5
```

```
##                           EVTYPE FATALITIES  INJURIES
## 766 TORNADOES, TSTM WIND, HAIL  25.000000  0.000000
## 62                COLD AND SNOW  14.000000  0.000000
## 775      TROPICAL STORM GORDON   8.000000 43.000000
## 519      RECORD/EXCESSIVE HEAT   5.666667  0.000000
## 127               EXTREME HEAT   4.363636  7.045455
```

- **Types of Events vs. Economic Consequences**

We can use aggregated() function to find the average property and crop damages by types of events, and store them into two new data frame. Combining these two new data frames grants us to reorder the data frame. Therefore, we can extract the top 5 types of events with highest property, crop, and total damages on average. The extracted top-5 data frames are intended to be plotted in Section 4.

```
## find the average property damage by types of events
storm.PROPDMG <- aggregate(PROPDMG ~ EVTYPE, storm.clean, mean)

## find the average crop damage by types of events
storm.CROPDMG <- aggregate(CROPDMG ~ EVTYPE, storm.clean, mean)

## combine two data frame
storm.economic <- cbind(storm.PROPDMG, storm.CROPDMG$CROPDMG)
colnames(storm.economic)[3] <- "CROPDMG"
storm.economic$TOTALDMG <- storm.economic$PROPDMG + storm.economic$CROPDMG

## extract the top 5 type of events with highest average property and crop damages, seperately and toge
storm.PROPDMG.top5 <- head(storm.PROPDMG[order(storm.PROPDMG$PROPDMG, decreasing = TRUE), ], 5)
```

```r
storm.CROPDMG.top5 <- head(storm.CROPDMG[order(storm.CROPDMG$CROPDMG, decreasing = TRUE), ], 5)
storm.economic.top5 <- head(storm.economic[order(storm.economic$TOTALDMG, decreasing = TRUE), ], 5)

## display the top 5 type of events with highest average property and crop damages, seperately and toge
storm.PROPDMG.top5
```

```
##                     EVTYPE PROPDMG
## 48         COASTAL EROSION     766
## 255   HEAVY RAIN AND FLOOD     600
## 528 RIVER AND STREAM FLOOD     600
## 36   BLIZZARD/WINTER STORM     500
## 143           FLASH FLOOD/     500
```

```r
storm.CROPDMG.top5
```

```
##                    EVTYPE CROPDMG
## 106 DUST STORM/HIGH WINDS     500
## 173          FOREST FIRES     500
## 775 TROPICAL STORM GORDON     500
## 353       HIGH WINDS/COLD     401
## 367       HURRICANE FELIX     250
```

```r
storm.economic.top5
```

```
##                     EVTYPE PROPDMG CROPDMG TOTALDMG
## 775  TROPICAL STORM GORDON     500     500     1000
## 48         COASTAL EROSION     766       0      766
## 255   HEAVY RAIN AND FLOOD     600       0      600
## 528 RIVER AND STREAM FLOOD     600       0      600
## 106  DUST STORM/HIGH WINDS      50     500      550
```

Section 4: Results

- **Types of Events vs. Population Health**

The data frames that store the top 5 types of events with highest average fatalities and injuries are already extracted in Section 3. We can plot them, using barplot in descending order.

```r
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```r
require(gridExtra)
```

```
## Loading required package: gridExtra
```

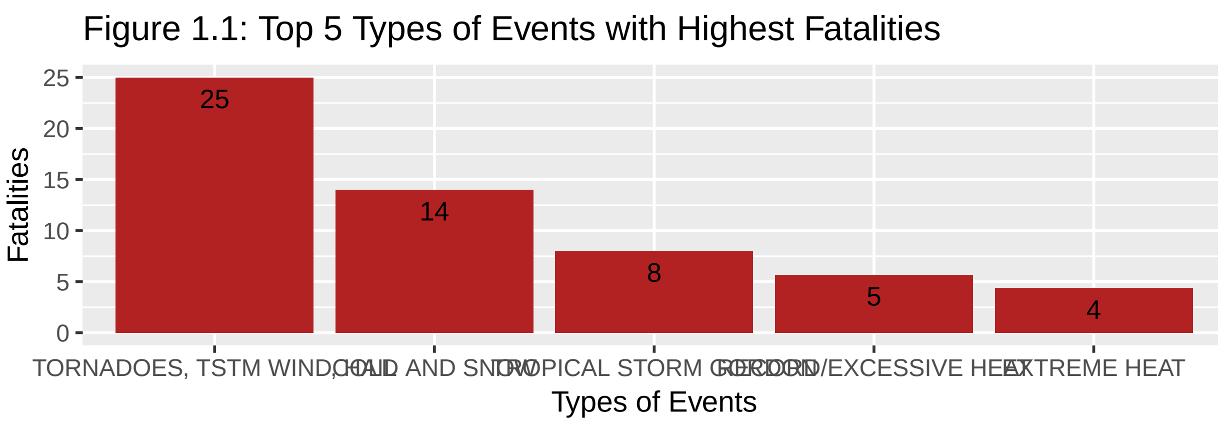
```r
## reorder
storm.FATALITIES.top5 <- transform(storm.FATALITIES.top5, EVTYPE = reorder(EVTYPE, -FATALITIES))
storm.INJURIES.top5 <- transform(storm.INJURIES.top5, EVTYPE = reorder(EVTYPE, -INJURIES))

plot1.1 <- ggplot(data = storm.FATALITIES.top5, aes(x = EVTYPE, y = FATALITIES)) +
    geom_bar(stat = "identity", fill = "firebrick") +
    geom_text(aes(label = as.integer(FATALITIES)), vjust = 1.6, color = "black", size = 3.5) +
    xlab("Types of Events") + ylab("Fatalities") +
    ggtitle("Figure 1.1: Top 5 Types of Events with Highest Fatalities")

plot1.2 <- ggplot(data = storm.INJURIES.top5, aes(x = EVTYPE, y = INJURIES)) +
```

```r
    geom_bar(stat = "identity", fill = "orange3") +
    geom_text(aes(label = as.integer(INJURIES)), vjust = 1.6, color = "black", size = 3.5) +
    xlab("Types of Events") + ylab("Injuries") +
    ggtitle("Figure 1.2: Top 5 Types of Events with Highest Injuries")

grid.arrange(plot1.1, plot1.2, ncol=1)
```



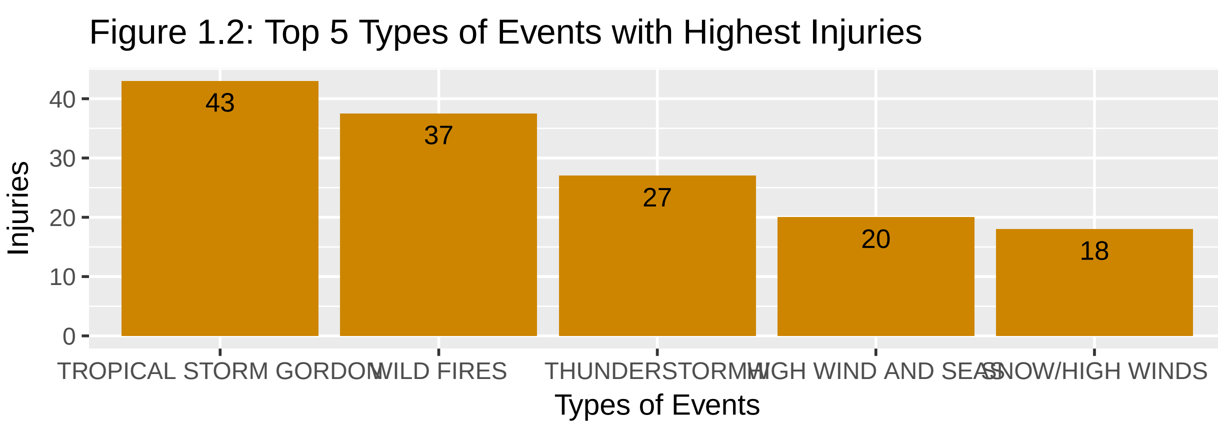


Figure 1.1 shows that **TORNADOES, TSTM WIND, HAIL** has the highest average fatalities across all types of events; Figure 1.2 shows that **TROPICAL STORM GORDON** has the highest average injuries across all types of events.

- **Types of Events vs. Economic Consequences**

The data frames that store the top 5 types of events with highest property, crop, and total damages on average are already extracted in Section 3. We can plot them, using barplot in descending order.

```r
require(ggplot2)
require(gridExtra)

## reorder
storm.PROPDMG.top5 <- transform(storm.PROPDMG.top5, EVTYPE = reorder(EVTYPE, -PROPDMG))
storm.CROPDMG.top5 <- transform(storm.CROPDMG.top5, EVTYPE = reorder(EVTYPE, -CROPDMG))
storm.economic.top5 <- transform(storm.economic.top5, EVTYPE = reorder(EVTYPE, -TOTALDMG))

plot2.1 <- ggplot(data = storm.PROPDMG.top5, aes(x = EVTYPE, y = PROPDMG)) +
    geom_bar(stat = "identity", fill = "olivedrab") +
    geom_text(aes(label = as.integer(PROPDMG)), vjust = 1.6, color = "black", size = 3.5) +
    xlab("Types of Events") + ylab("Property Damage") +
```

```
    ggtitle("Figure 2.1: Top 5 Types of Events with Highest Property Damage")

plot2.2 <- ggplot(data = storm.CROPDMG.top5, aes(x = EVTYPE, y = CROPDMG)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    geom_text(aes(label = as.integer(CROPDMG)), vjust = 1.6, color = "black", size = 3.5) +
    xlab("Types of Events") + ylab("Crop Damage") +
    ggtitle("Figure 2.2: Top 5 Types of Events with Highest Crop Damages")

plot2.3 <- ggplot(data = storm.economic.top5, aes(x = EVTYPE, y = TOTALDMG)) +
    geom_bar(stat = "identity", fill = "blueviolet") +
    geom_text(aes(label = as.integer(TOTALDMG)), vjust = 1.6, color = "black", size = 3.5) +
    xlab("Types of Events") + ylab("Total Damages") +
    ggtitle("Figure 2.3: Top 5 Types of Events with Highest Total Damages")

grid.arrange(plot2.1, plot2.2, plot2.3, nrow=3)
```

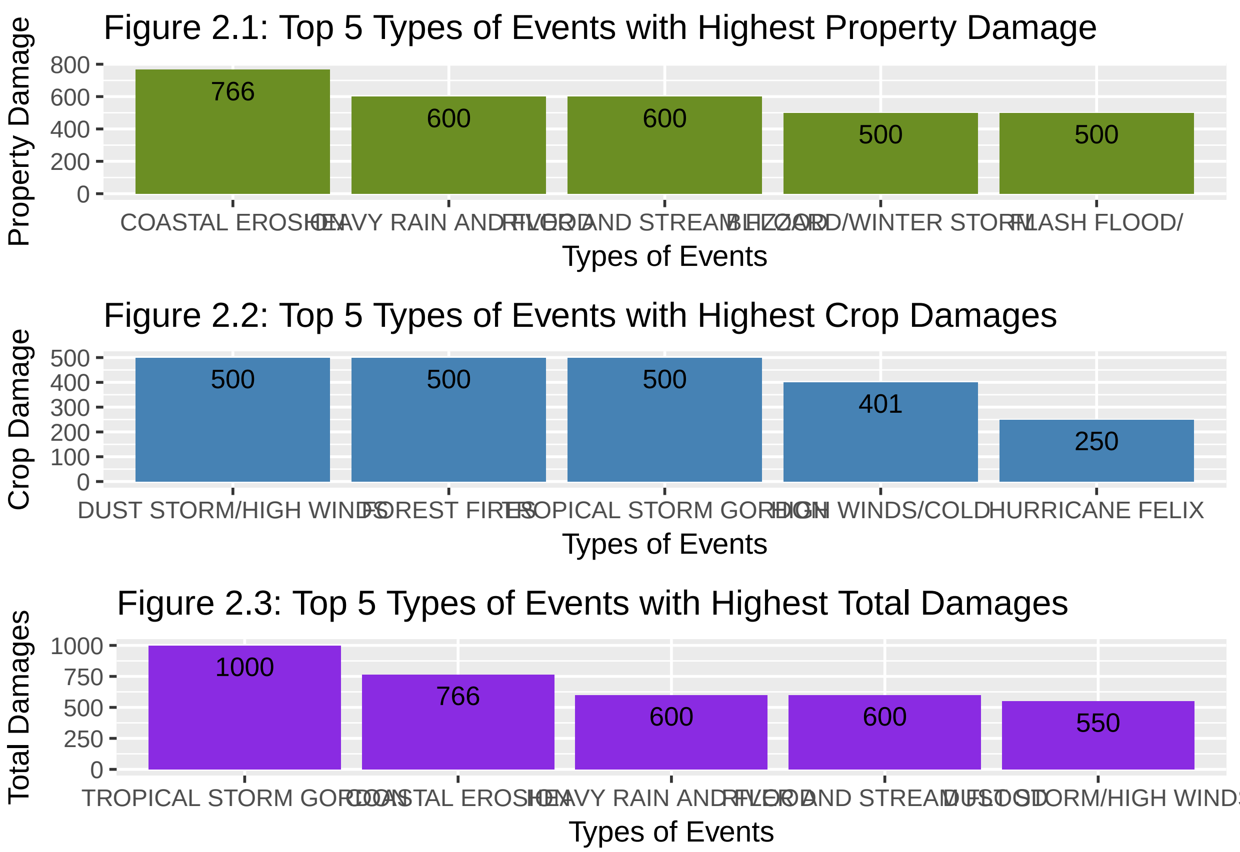


Figure 2.1 shows that **COASTAL EROSION** has the highest average property damage across all types of events; Figure 2.2 shows that **DUST STORM/HIGH WINDS** has the highest average crop damage across all type of events; Figure 2.3 shows that **TROPICAL STORM GORDON** has the highest total damages on average.