Reproducible Research: Peer Assessment 2

NOAA Storm Data Impact Analysis

This document contains an analysis of NOAA storm data to be submitted for the Reproducible Research class offered on Coursera. Specifically, we are looking to answer the following questions.

1. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

Synopsis

Overall, we found that TORNADOes had both the greatest health consequences and property damage. Looking at the data, each individual TORNADO event is devastating and they occur in enough frequency to accumulate a large total.

Data Processing

First things first: reading in the raw data and looking at some of the relevant information.

```
stormdata <- read.csv('stormdata.csv.bz2')
names(stormdata)
```

```
##  [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"
```

We can cut out some of the columns that won't be relevant to our analysis. To answer the question about population health, we should save the columns indicating the fatalities and injuries sustained during the event. To answer the question about the economic consequences, we should save the columns indicating both property and crop damage, in US dollars.

```
stormdata <- stormdata[, c('EVTYPE', 'FATALITIES', 'INJURIES', 'PROPDMG', 'CROPDMG')]
```

We note also that the capitalization of the event type is not standardized. For example, both "BLACK ICE" and "Black Ice" show up as event types. We can solve this by changing all event types to uppercase.

```
stormdata$EVTYPE <- sapply(stormdata$EVTYPE, toupper)
```

With all the relevant information in hand, we can proceed to the analysis.

Analysis

For the analysis, we'd like to know which events are the deadliest and cause the most economic damage on an absolute and a per-event basis. A particular type of event might be very deadly, but it also may not happen that often. For this reason, we would like to know both the average and total.

```
library('dplyr')
summ <- stormdata %>% group_by_(.dots=as.symbol('EVTYPE')) %>% summarize_each(funs(mean, sum))
```

Results

Health Consequences

First we sort the dataframe by the total and average number of fatalities and injuries and looking at the top entry in each case.

```
summ[with(summ, order(-FATALITIES_sum))[1], ]
```

```
## # A tibble: 1 × 9
##    EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean CROPDMG_mean
##     <chr>           <dbl>         <dbl>        <dbl>        <dbl>
## 1 TORNADO       0.0928741      1.506067     52.96211     1.649056
## # ... with 4 more variables: FATALITIES_sum <dbl>, INJURIES_sum <dbl>,
## #   PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

```
summ[with(summ, order(-FATALITIES_mean))[1], ]
```

```
## # A tibble: 1 × 9
##                       EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean
##                        <chr>           <dbl>         <dbl>        <dbl>
## 1 TORNADOES, TSTM WIND, HAIL              25             0          1.6
## # ... with 5 more variables: CROPDMG_mean <dbl>, FATALITIES_sum <dbl>,
## #   INJURIES_sum <dbl>, PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

```
summ[with(summ, order(-INJURIES_sum))[1], ]
```

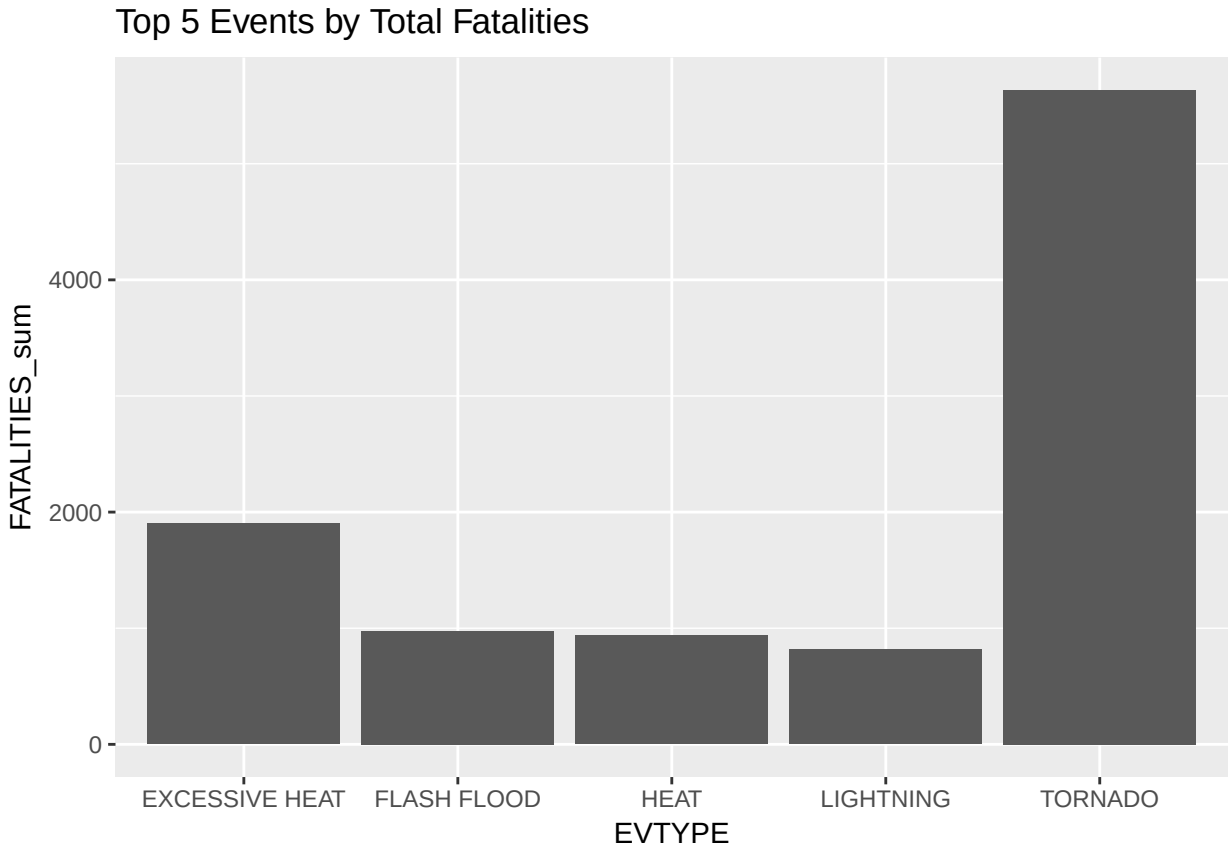
```
## # A tibble: 1 × 9
##    EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean CROPDMG_mean
##     <chr>           <dbl>         <dbl>        <dbl>        <dbl>
## 1 TORNADO       0.0928741      1.506067     52.96211     1.649056
## # ... with 4 more variables: FATALITIES_sum <dbl>, INJURIES_sum <dbl>,
## #   PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

```
summ[with(summ, order(-INJURIES_mean))[1], ]
```

```
## # A tibble: 1 × 9
##                   EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean
##                    <chr>           <dbl>         <dbl>        <dbl>
## 1 TROPICAL STORM GORDON               8            43          500
## # ... with 5 more variables: CROPDMG_mean <dbl>, FATALITIES_sum <dbl>,
## #   INJURIES_sum <dbl>, PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

Overall, it seems that TORNADOes are the deadliest weather events. We can visualize this information by plotting the top 5 entries by total fatalities.

```
library('ggplot2')
by_total_fatalities <- head(summ[with(summ, order(-FATALITIES_sum)), ], 5)
ggplot(data=by_total_fatalities, aes(x=EVTYPE, y=FATALITIES_sum)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle('Top 5 Events by Total Fatalities')
```



Economic Damage

We'd also like to summarize the economic consequences of each event type. We take a similar approach as for health consequences.

```
summ[with(summ, order(-PROPDMG_sum))[1], ]
```

```
## # A tibble: 1 × 9
##    EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean CROPDMG_mean
##     <chr>           <dbl>         <dbl>        <dbl>        <dbl>
## 1 TORNADO       0.0928741      1.506067     52.96211     1.649056
## # ... with 4 more variables: FATALITIES_sum <dbl>, INJURIES_sum <dbl>,
## #   PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

```
summ[with(summ, order(-PROPDMG_mean))[1], ]
```

```
## # A tibble: 1 × 9
##            EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean CROPDMG_mean
##             <chr>           <dbl>         <dbl>        <dbl>        <dbl>
## 1 COASTAL EROSION               0             0          766            0
## # ... with 4 more variables: FATALITIES_sum <dbl>, INJURIES_sum <dbl>,
## #   PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

```
summ[with(summ, order(-CROPDMG_sum))[1], ]
```

```
## # A tibble: 1 × 9
##   EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean CROPDMG_mean
##    <chr>           <dbl>         <dbl>        <dbl>        <dbl>
```

```
## 1   HAIL    5.196407e-05   0.004714873     2.385821     2.007879
## # ... with 4 more variables: FATALITIES_sum <dbl>, INJURIES_sum <dbl>,
## #   PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

```
summ[with(summ, order(-CROPDMG_mean))[1], ]
```

```
## # A tibble: 1 × 9
##                   EVTYPE FATALITIES_mean INJURIES_mean PROPDMG_mean
##                    <chr>           <dbl>         <dbl>        <dbl>
## 1 DUST STORM/HIGH WINDS               0             0           50
## # ... with 5 more variables: CROPDMG_mean <dbl>, FATALITIES_sum <dbl>,
## #   INJURIES_sum <dbl>, PROPDMG_sum <dbl>, CROPDMG_sum <dbl>
```

Here, the answers are a little less clear as we get a different top answer for each. Overall, I have to give the prize to TORNADOes again as they caused the most overall damage in total magnitude. We can visualize this information as well.

```
library('reshape2')
by_total_propdmg <- head(summ[with(summ, order(-PROPDMG_sum)), ], 5)
melted <- melt(by_total_propdmg, c('PROPDMG_sum', 'CROPDMG_sum'), id.vars='EVTYPE')
ggplot(data=melted, aes(x=EVTYPE, y=value, fill=variable)) +
    geom_bar(stat="identity", position=position_dodge()) +
    ggtitle('Top 5 Events by Total Property Damage') +
    ylab('Damage (dollars)')
```