Project 2 - Reproducible Research

Rich Huebner

January 25, 2016

Introduction & Synopsis

The basic goal of this assignment was to explore the NOAA Storm Database and answer some basic questions about severe weather events. Data is from the National Weather Service storm data, and is available here. The data contains the occurrence of storms and other significant weather phenomena having sufficient intensity to cause loss of life, injuries, property damage, etc. The analysis uses the Storm database to answer the questions listed below. All code and processing steps are shown below, along with the results. The data is processed to standarize/normalize some of the values.

The data had to be preprocessed prior to conducting the actual analysis, and prior to determining the answers to the questions. Preprocessing involved ensuring that NAs (no value) were converted to zeros. Additionally, the data had to be aggregated by the event type (EVTYPE) as well, so there are functions in the preprocessing that do that as well.

Libraries Used in this Analysis The following libraries are used in the analysis.

ggplot2 for graphing and plotting
dplyr for subsetting and data manipulation

Data Set Information Detailed Documentation here: https://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf

Data Set here: https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2

The data set file, when extracted, is about 500Mb. Loading the file upon first running the script takes time, so please be patient.

Data Analysis and Questions in this Study The data analysis addressed the following questions:

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

Data Processing

Load the required libraries for R.

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
setwd("D:\\data\\RProjects\\RResearch_Project2")
data <- read.csv("repdata-data-StormData.csv", header=TRUE, sep=",")

head(data)
```

Load the data set into R.

```
##   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
```

Clean the data by normalizing the data set.

```
#### First, I will only use the variables I need in this analysis. So, create a subset.
scols = c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG", "CROPDMGEXP")
subdata <- data[scols]
```

```
#### Fill in 0s for anything that's missing in the dataset
subdata[subdata$FATALITIES == ""] <- 0
subdata[subdata$INJURIES == ""] <- 0
subdata[subdata$PROPDMG == ""] <- 0
subdata[subdata$CROPDMG == ""] <- 0

head(subdata)
```

Update event names – essentially normalizing them, "wind, WiNd" = "WIND"

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

```
#### Clean the data by normalizing the data set. "WInd, wind" = "WIND"

subdata$PROPDMGEXP[subdata$PROPDMGEXP == ""] <- 0
subdata$EVTYPE <- gsub("^HEAT$", "EXCESSIVE HEAT", subdata$EVTYPE )
subdata$EVTYPE <- gsub("^TSTM WIND$", "THUNDERSTORM WIND", subdata$EVTYPE)
subdata$EVTYPE <- gsub("^THUNDERSTORM WIND$", "THUNDERSTORM WIND", subdata$EVTYPE)
```

```
f <- aggregate(subdata$FATALITIES, by=list(subdata$EVTYPE), sum, na.rm = TRUE)
names(f) <- c("Category", "Total")
fsort <- f[order(-f$Total), ]
topf <- fsort[1:10, ]
topf$Category <- factor(topf$Category, levels=topf$Category, ordered=TRUE)
```

Get aggregated data on fatalities.

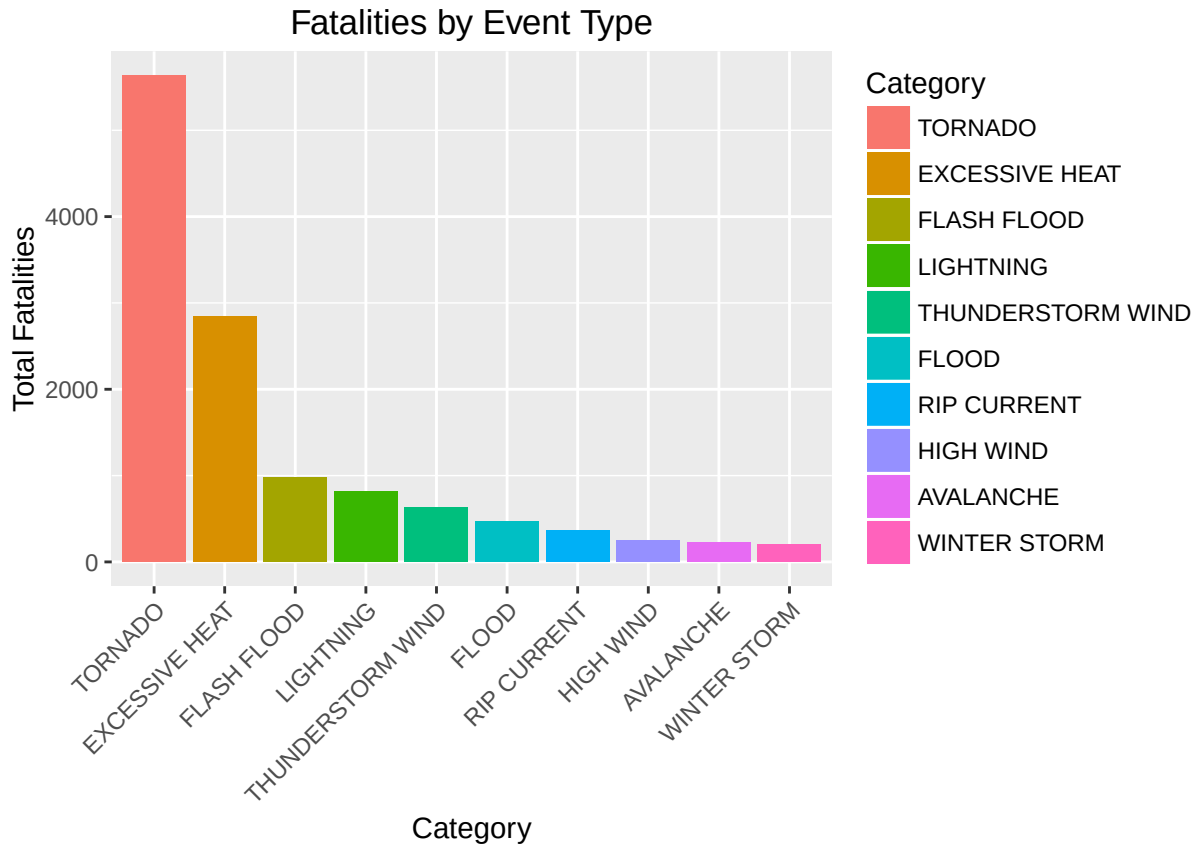
```
e <- aggregate(subdata$PROPDMG, by=list(subdata$EVTYPE), sum, na.rm=TRUE)
names(e) <- c("Category", "Total")
esort <- e[order(-e$Total), ]
tope <- esort[1:10, ]
tope$Category <- factor(tope$Category, levels=tope$Category, ordered=TRUE)
```

Get aggregated data on economic consequences

Results

Q 1. Across the U.S., which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health? Here's a plot of the data to answer question #1.

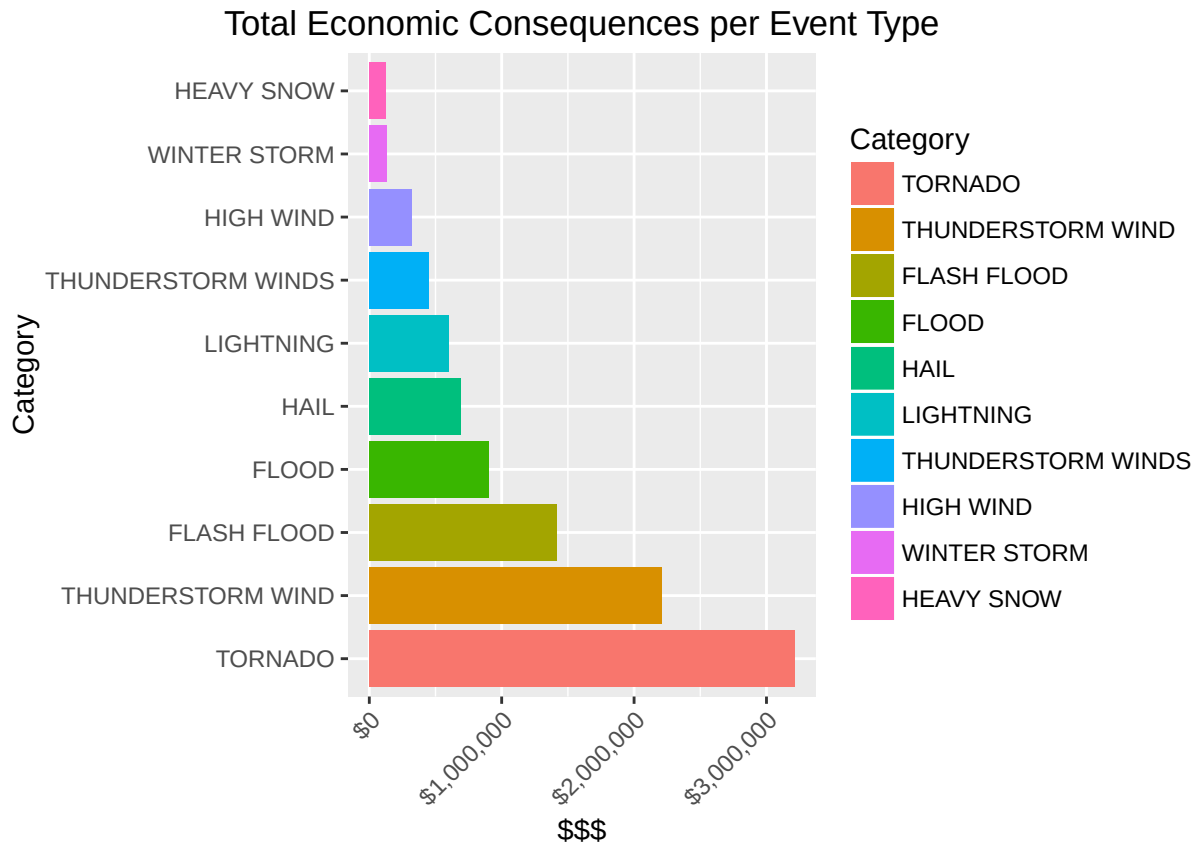
```
gplot <- ggplot(topf, aes(x=Category, y=Total, fill=Category)) +
          geom_bar(stat="identity") +
          labs(x="Category", y="Total Fatalities", title="Fatalities by Event Type") +
          theme(axis.text.x = element_text(angle=45, hjust=1))
print(gplot)
```



Solution: Based on the graph, the MOST harmful with respect to population health is the TORNADO, followed by EXCESSIVE HEAT and FLASH FLOODING.

Q 2. Across the U.S., which types of events have the greatest economic consequences? Here's the second plot. This one addresses question #2.

```
gplot <- ggplot(tope, aes(x=Category, y=Total, fill=Category)) +
     geom_bar(stat="identity") +
     coord_flip() +
     labs(x="Category", y="$$$", title="Total Economic Consequences per Event Type") +
     theme(axis.text.x = element_text(angle=45, hjust=1)) +
     scale_y_continuous(labels = scales::dollar)
print(gplot)
```



In terms of greatest economic impact, the TORNADO is the highest, followed by THUNDERSTORM WINDS and FLASH FLOODING.