Storm event Type Analysis: RepResearch Proj 2

Synopsis

In a brief analysis of the storm data provided a clear conclusion can be seen that while there are many storm events that can impact us, a bulk of the harm of cost to the population occur in a few main events.

Data Processing

Reading Import

The read.csv command will handle the .bz2 extension. I also created a txt file of the 48 weather event types and made that into a data frame.

```
storm <- read.csv("C:/Users/563571/Desktop/Tech Tank/Coursera/RepResearch/repdata-data-St
ormData.csv.bz2")
evtype48 <- read.csv("C:/Users/563571/Desktop/Tech Tank/Coursera/RepResearch/evtype.txt",
header=FALSE)
evtype48 = as.data.frame(evtype48)
```

Data Formatting

This anlysis is focusing on weather event types. In the initial data, there are over 900 types, while the NOAA lists 48 in their explanation on the data. What I tried to accomplish was to standarizde the event types and eliminate or combine data that was poorly entered.

I manually created a text file of the 48 events. The process below uses the grep function in R to find matches, count the number of matches (just to see how it is matching), rename any event matches, and remove rows

```
# evtype48$V1 is the set of 48 event type from the storm data documentation
# Sets the new EVTYPE to match one from the 48, if possible
for (n in evtype48$V1) {
  storm$EVTYPE[grep(n,storm$EVTYPE,ignore.case=TRUE)] = n
}

# Keeps only the rows which have a event type match.
storm2 = storm[which(storm$EVTYPE%in%evtype48$V1),]
```

Results in which your results are presented.

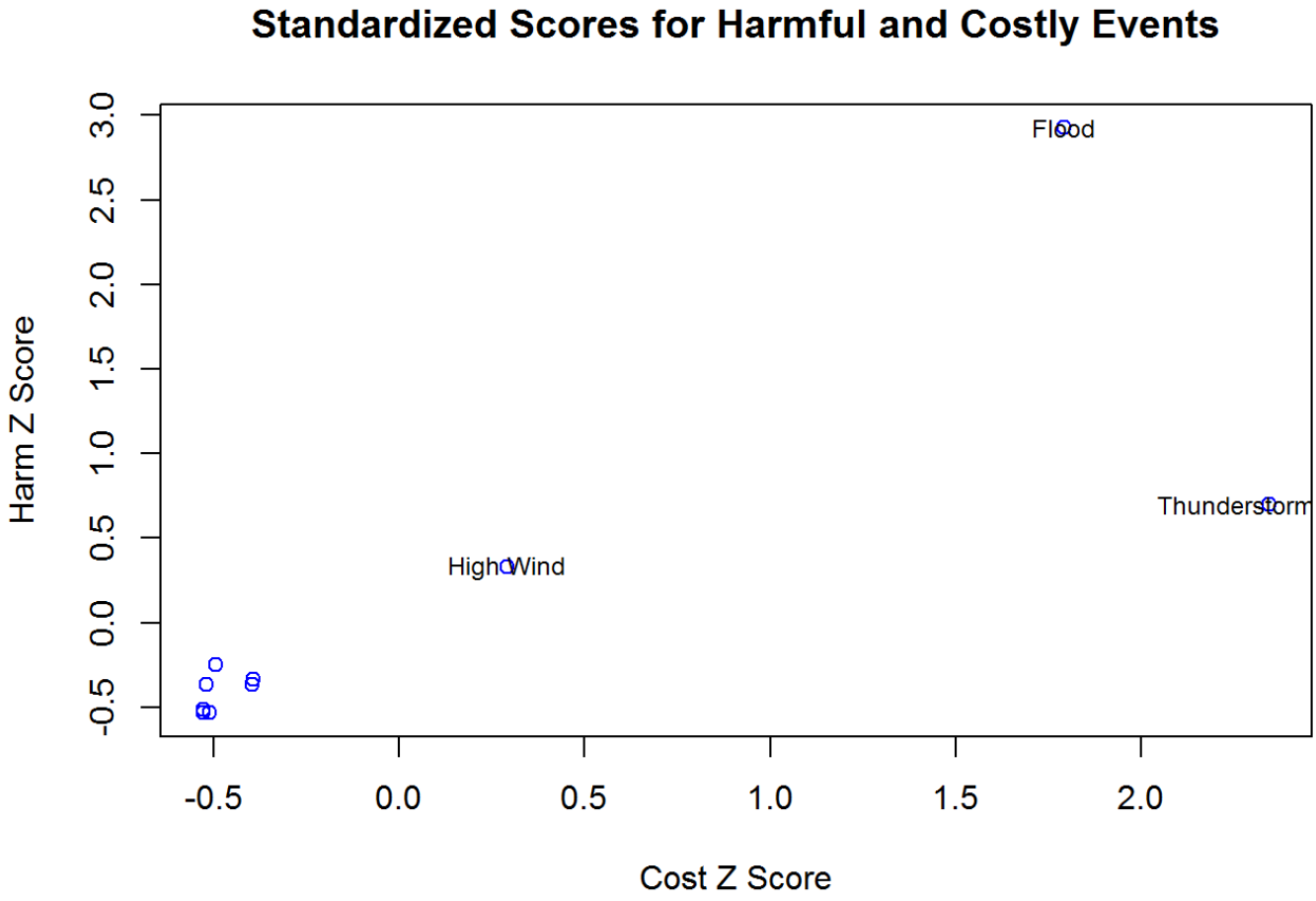
I interpreted “population harm” as injuries + fatalities so keep analysis simple. Similarly, economic costs were considered property damage + crop damage. I also standardized by subtracting the means and dividing by the standard deviations for cost and harm, respectively.

The below code shows how I constructed this data.

```
costly = aggregate(PROPDMG+CROPDMG~EVTYPE,FUN = sum,data = storm2)
harmful = aggregate(FATALITIES+INJURIES~EVTYPE,FUN = sum,data = storm2)
costly$z = (costly[,2]-mean(costly[,2]))/sd(costly[,2])
harmful$z = (harmful[,2]-mean(harmful[,2]))/sd(harmful[,2])
data = merge(costly,harmful,by="EVTYPE")
names(data)=c("EVTYPE","COSTS","z.costs","HARM","z.harm")
```

To see which events are high in both cost and harm when compared to all events, we can create a scatterplot if the Z scores.

```
plot(data$z.costs,data$z.harm,main="Standardized Scores for Harmful and Costly Events",xl
ab="Cost Z Score",ylab="Harm Z Score",col = "blue")
text(x = data$z.costs[which(data$z.costs>.25 | data$z.harm>.25)], y = data$z.harm[which(d
ata$z.costs>.25 | data$z.harm>.25)],labels = data$EVTYPE[which(data$z.costs>.25 | dat
a$z.harm>.25)],cex=.75)
```



For simplicity on the plot, the only points labeled are ones where a z score is above 0.25. These event types were:

```
data$EVTYPE[which(data$z.costs>.25 | data$z.harm>.25)]
```

```
## [1] Flood             High Wind         Thunderstorm Wind
## 985 Levels:    HIGH SURF ADVISORY  COASTAL FLOOD ... WND
```

We see that high wind, floods, and thunderstorms are each very high with respect to cost and harm relative to other weather event types.