Analysis of Storm DataSet

The purpose of this analysis is to answer two key questions: Which weather events were most harmful to the population health, and which weather events had the greatest economic consequences. Harm to population health is measured in terms of injuries and fatalities, while economic consequences are measured in terms of dollar amounts. In order to come up with reasonable conclusions, a certain amount of cleaning and preparation of the data had to first be performed, which will also be included in this report. The majority of preparation involved restricting our scope to just certain columns, analyzing only those weather events that occured in 1996 or more recently, and creating a few new columns in order to help with analysis later on.

DATA PROCESSING

The data was obtained through a link provided within the Coursera course called Reproducible Research (June, 2018). The unzipped file is approximately 500mb.

```
library(dplyr)
```

```
##
## 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
```

```
library(stringdist)
library(ggplot2)
```

```
#Read the csv file into Data Frame
Storm <- read.csv("repdata_data_StormData.csv.bz2")
```

Create a dictionary of 48 official event types.The 900+ event types in the original data set will be mapped to one of these 48 official event types using the amatch() function in the “stringdist” pachage, using the Levenshtein distance to compare the similarities and differences between strings.

```
#Created a dictionary
dictionary <- c("Astronomical Low Tide","Avalanche","Blizzard","Coastal Flood","Col
d/Wind Chill","Debris Flow","Dense Fog","Dense Smoke","Drought","Dust Devil","Dust
Storm","Excessive Heat","Extreme Cold/Wind Chill","Flash Flood","Flood","Freezing F
og","Frost/Freeze","Funnel Cloud","Hail","Heat","Heavy Rain","Heavy Snow","High Sur
f","High Wind","Hurricane/Typhoon","Ice Storm","Lakeshore Flood","Lake-Effect Sno
w","Lightning","Marine Hail","Marine High Wind","Marine Strong Wind","Marine Thunde
rstorm Wind","Rip Current","Seiche","Sleet","Storm Tide","Strong Wind","Thunderstor
m Wind","Tornado","Tropical Depression","Tropical Storm","Tsunami","Volcanic As
h","Waterspout","Wildfire","Winter Storm","Winter Weather")
#Make both the dictionary and Storm$EVTYPE to lowercase for more accurate result.
dictionary <- tolower(dictionary)
Storm$EVTYPE <- tolower(Storm$EVTYPE)
Storm$EVTYPE <- dictionary[amatch(Storm$EVTYPE,dictionary,method = "lv",maxDist = 2
0)]
```

Create a data Frame with 2 columns representing Casualties.

```
Pop_health <- Storm %>% group_by(EVTYPE) %>% summarise(Casualties = sum(10*FATALITI
ES + INJURIES))
```

Further subset the database as said

```
## Removes all rows that have a 0 in each of the 2 columns PROPDMG, CROPDMG and  PR
OPDMGEXP,CROPDMGEXP with "k or B or M" strings.
## In other words, if a weather event led to no casualites or economic damage, we w
ill ignore it from this point on.
Storm <- Storm %>% filter(PROPDMGEXP %in% c("K","B","M") & CROPDMGEXP %in% c
("K","B","M") | CROPDMG != 0  | PROPDMG != 0)
```

The last step of processing the data is to find out what the total dollar amounts of property damage and crop damage are, as well as an overall total. This will require a bit of care because the amount of damage is found in one column as the base value, and in the adjacent column as an exponent. Luckily, at this point, the only values left in the exponents column are “K”, “M”, and “B”, whereas before we had done some subsetting, there were many nonstandard values.

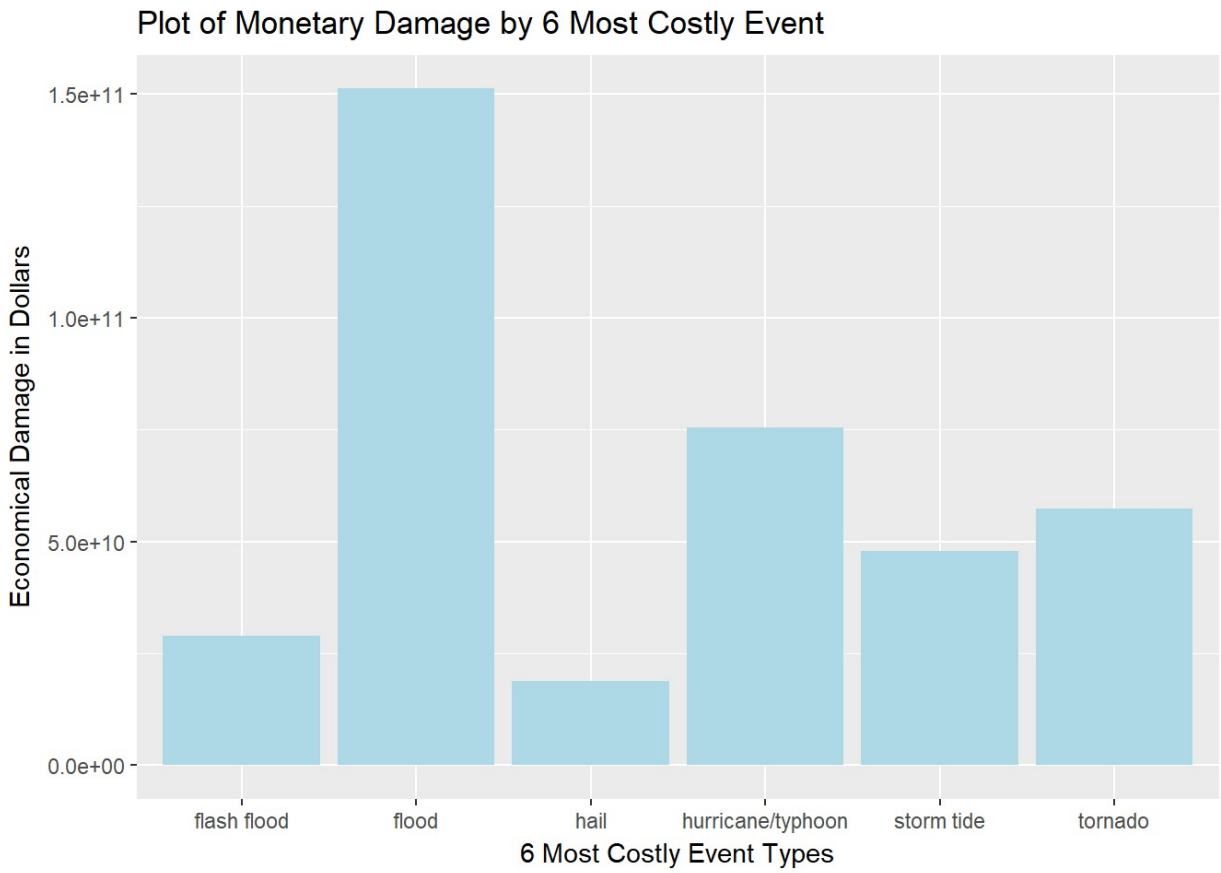
```
Storm[Storm$PROPDMGEXP == "M",25] <- Storm[Storm$PROPDMGEXP == "M",25]*10^6
Storm[Storm$PROPDMGEXP == "K",25] <- Storm[Storm$PROPDMGEXP == "K",25]*10^3
Storm[Storm$PROPDMGEXP == "B",25] <- Storm[Storm$PROPDMGEXP == "B",25]*10^9
Storm[Storm$CROPDMGEXP == "M",27] <- Storm[Storm$CROPDMGEXP == "M",27]*10^6
Storm[Storm$CROPDMGEXP == "K",27] <- Storm[Storm$CROPDMGEXP == "K",27]*10^3
Storm[Storm$CROPDMGEXP == "B",27] <- Storm[Storm$CROPDMGEXP == "B",27]*10^9

## One last step, let's create a data Frame with both Casualties and Economical dam
age
Economical_DMG <- Storm %>% group_by(EVTYPE) %>% summarise(TotalDMG = sum(PROPDMG
+ CROPDMG))
Total_DMG <- inner_join(Economical_DMG,Pop_health,by = "EVTYPE")
```

Now that we have totals for the amount of casualties that occured in each event, as well as how much economic damage occured in terms of property, crops, and in total, we can begin to pull some results from the data.

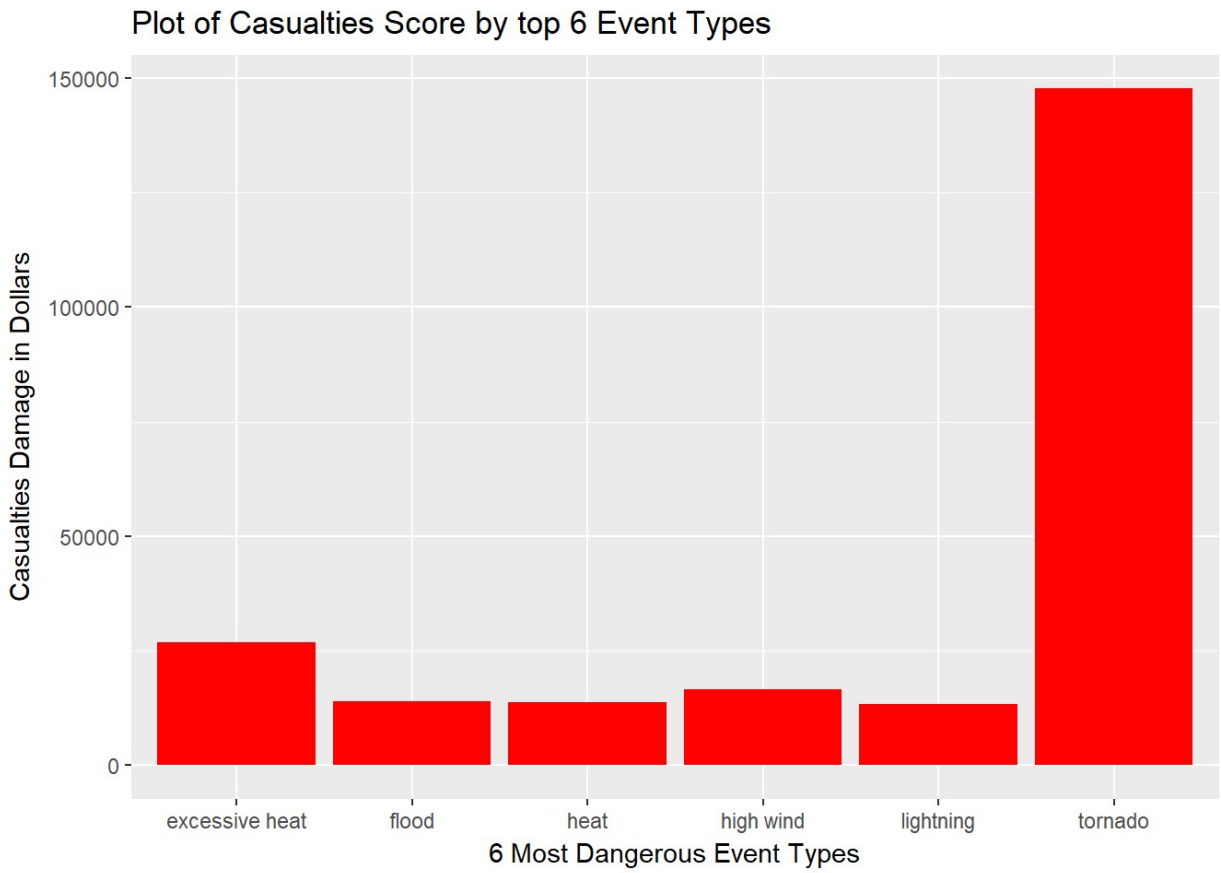
RESULTS

Again, our primary goal here is find out which weather events caused the most damage in terms of casualties, as well as in terms of economic damage.

```
#Here we plot the 6 most costly categories
Total_DMG <- Total_DMG %>% arrange(desc(TotalDMG))
PlotA <- ggplot(Total_DMG[1:6,],aes(x = EVTYPE,y = TotalDMG))
PlotA <- PlotA + geom_bar(fill = "light blue",stat = "identity") + xlab("6 Most Cos
tly Event Types") + ylab("Economical Damage in Dollars") + ggtitle("Plot of Monetar
y Damage by 6 Most Costly Event")
PlotA
```




```
#Here we Plot the Casualties Score by top 6 Events
Total_DMG <- Total_DMG %>% arrange(desc(Casualties))
PlotB <- ggplot(Total_DMG[1:6,],aes(x = EVTYPE,y = Casualties))
PlotB <- PlotB + geom_bar(fill = "red",stat = "identity") + xlab("6 Most Dangerous
Event Types") + ylab("Casualties Damage in Dollars") + ggtitle("Plot of Casualties
Score by top 6 Event Types")
PlotB
```


Plot of Casualties Score by top 6 Event Types
Casualties Damage in Dollars
150000
100000
50000
0
excessive heat
flood
heat
high wind
lightning
tornado
6 Most Dangerous Event Types