Economic and Health Impacts of Hazardous Weather Conditions

Synopsis

The basic goal of this assignment is to explore the NOAA Storm Database and answer some basic questions about severe weather events. This analysis considers the Storm Data provided by the National Weather Service. The dataset contains records of occurance of storms and other significant weather phenomena that had sufficient intensity to cause loss of life, injuries, significant property damage, and/or disruption of commerce. My data analysis finds that Tornado is the event most harmful to human health as it causes maximum number of injuries and fatalities. The analysis of the economic data shows that Tornado and Wind are the two events causing most damage to property and crops.

Data Processing

The first step in the data analysis is to load the data provided by the National Weather Service. The original data will be stored as rawdata and will never be changed.

```
#Load Packages
library(ggplot2)
library(data.table)
library(dplyr)

# Read data
rawdata <- read.csv(bzfile("repdata-data-StormData.csv.bz2"))
rawnrows <- nrow(rawdata)
rawncols <- ncol(rawdata)
```

The original dataset contains 902297 rows and 37 variables. Variable names are shown below:

```
print(names(rawdata))
```

```
##  [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"
```

In order to assess which events are most harmful with respect to population health, I am going to consider the following variables:

- Fatalities
- Injuries

In order to assess which events have the greatest economic consequences, I am going to consider the following variables:

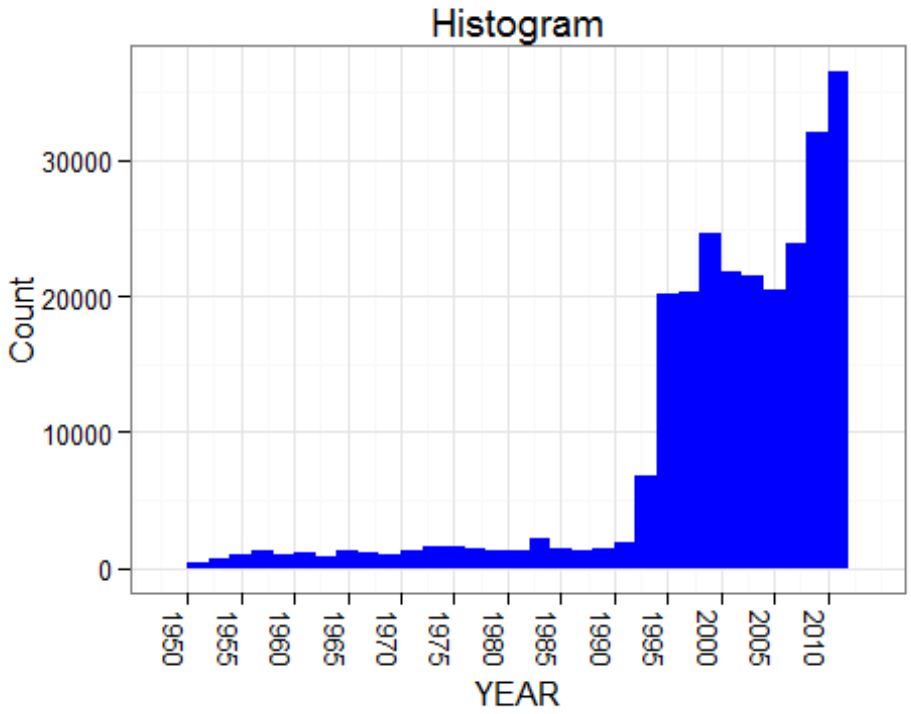
- Property Damage
- Crop Damage

Given that the purpose of the analysis is to consider those events that have drastic health or economic consequences, I am going to drop all observations where FATALITIES, INJURIES, PROPDMG and CROPDMG are zero. This will significantly shorten the dataset.

```
shortdata <- rawdata[rawdata$FATALITIES != 0 |
                         rawdata$INJURIES !=0|
                         rawdata$PROPDMG !=0|
                         rawdata$CROPDMG !=0,]
nrows <- nrow(shortdata)
```

The size of the dataset is reduced to 254633 rows. The second step, is to see how the event frequency is distributed across years. Since data collecting was less structured during earlier years, we may see that there are a lot more data in the recent years.

```
#Convert date to the actual date format
shortdata$BGN_DATE <- as.Date(shortdata$BGN_DATE, format = "%m/%d/%Y")
shortdata$YEAR <- year(shortdata$BGN_DATE)
#Let's create a histogram to see how observations change over years
histogram <- ggplot(data=shortdata, aes(x=YEAR))+
    geom_histogram(fill="blue", binwidth = 2)+
    labs(title = "Histogram", x= "YEAR", y = "Count")+
    scale_x_continuous(breaks = seq(1950, 2011,5))+
    theme_bw()+
    theme(axis.text.x=element_text(angle=-90))
print(histogram)
```



Based on the graph above, we conclude that beginning in 1995, there was a dramatic increase in the amount of collected data. Therefore, we can change our dataset sample to begin in 1995.

```
shortdata <- shortdata[shortdata$YEAR > 1994,]
uniquevals <- length(unique(shortdata$EVTYPE))
nrows <- nrow(shortdata)
```

After reducing the sample size, the size of the dataset is reduced to 211775 rows.

There are 372 unique weather conditions in the dataset. However, I need to find an approach to categorize them into fewer categories. Frist, I will clean up the names so that we can have fewer weather categories to analyze. I will remove trailing spaces and common typos as well as reduce the total numbe of unique categories.

```
#Clean Up the Names in EVTYPE
shortdata$EVTYPE <- gsub("WINDS", "WIND", shortdata$EVTYPE, ignore.case=TRUE)
shortdata$EVTYPE <- gsub("RAINS", "RAIN", shortdata$EVTYPE, ignore.case=TRUE)
shortdata$EVTYPE <- gsub("THUNDERSTORMS", "THUNDERSTORM", shortdata$EVTYPE,
ignore.case=TRUE)
shortdata$EVTYPE <- gsub("FLOODING", "FLOOD", shortdata$EVTYPE,
ignore.case=TRUE)
shortdata$EVTYPE <- gsub("ADVISORY", "", shortdata$EVTYPE, ignore.case=TRUE)
shortdata$EVTYPE <- gsub("TSTM", "", shortdata$EVTYPE, ignore.case=TRUE)
shortdata$EVTYPE <- gsub("FREEZING", "FREEZE", shortdata$EVTYPE,
ignore.case=TRUE)
shortdata$EVTYPE <- gsub("ICY", "ICE", shortdata$EVTYPE, ignore.case=TRUE)
```

```
shortdata$EVTYPE <- gsub("WILDFIRE", "FIRE", shortdata$EVTYPE,
ignore.case=TRUE)
#Simplify Categories
weathercat = c("LIGHTNING", "FREEZE", "FLOOD", "TSUNAMI", "TORNADO",
"BLIZZARD",
               "DROUGHT", "HAIL", "HEAT", "SLEET", "FIRE", "COLD", "WIND",
               "FOG", "STORM", "RAIN", "SNOW", "HURRICANE", "THUNDERSTORM",
               "SEICHE", "ICE", "DUST", "SMOKE", "SURF", "AVALANCHE", "RIP
CURRENTS",
               "MUDSLIDE", "TIDE", "VOLCANIC ASH", "FUNNEL CLOUD")
for(elem in weathercat) {
temp <- which(grepl(elem, shortdata$EVTYPE, ignore.case=TRUE))
    for (j in temp) {
        shortdata$EVTYPE[j] = elem
    }
}
shortdata$EVTYPE <- gsub("(^[[:space:]]+|[[:space:]]+$)","",shortdata$EVTYPE)
#Remove anything that is in parenthesis
shortdata$EVTYPE <- gsub(" *\\(.*?\\)* ", "", shortdata$EVTYPE)
#Everything that isn't one of the categories above, will be classified as
OTHER
for (i in 1:nrow(shortdata)){
    if(is.element(shortdata$EVTYPE[i], weathercat) == FALSE){
        shortdata$EVTYPE[i] <- "OTHER"
    }
}

uniquevals <- length(unique(shortdata$EVTYPE))
othercount <- nrow(shortdata[shortdata$EVTYPE == "OTHER",])
allobs <- nrow(shortdata)
othershare <- (othercount/allobs)*100
```

Based on the documentation and the categories in the table, weather types have been reduced to 30 categories. Those categories are as follows:

```
print(unique(shortdata$EVTYPE))
```

```
##  [1] "WIND"          "HURRICANE"     "FOG"           "HAIL"
##  [5] "OTHER"         "STORM"         "LIGHTNING"     "FLOOD"
##  [9] "TORNADO"       "HEAT"          "RAIN"          "COLD"
## [13] "AVALANCHE"     "SNOW"          "SLEET"         "DUST"
## [17] "SURF"          "FIRE"          "FREEZE"        "ICE"
## [21] "DROUGHT"       "BLIZZARD"      "RIP CURRENTS"  "MUDSLIDE"
## [25] "FUNNEL CLOUD"  "SEICHE"        "VOLCANIC ASH"  "TIDE"
## [29] "TSUNAMI"       "SMOKE"
```

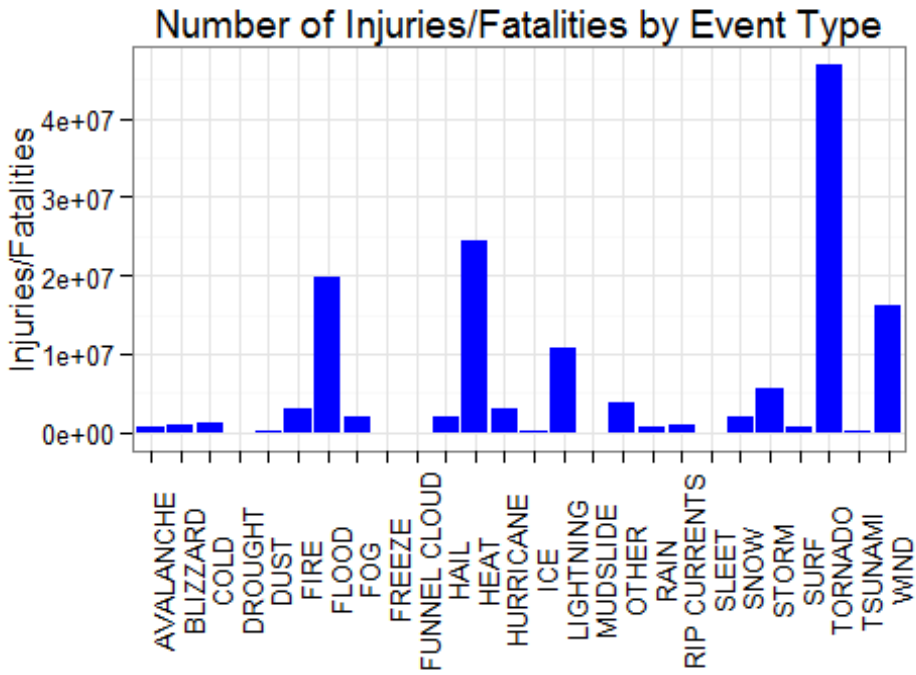
OTHER category includes 2129 observations, while total number of observations is 211775. Therefore, OTHER accounts for 1.01% share of all observations.

Results

Now that the data processing is complete, we can take a look at the results.

Let's begin by looking at the events that cause maximum health damage. For that, let's look at ALL INCIDENTS, which are defined as the sum of FATALITIES and INJURIES.

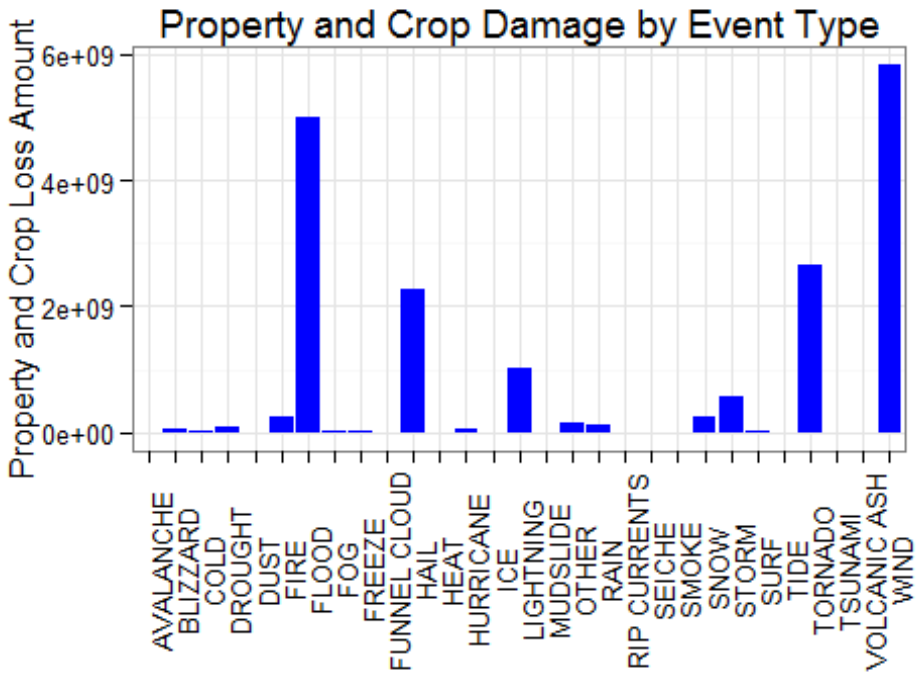
```
healthdata <- shortdata[shortdata$FATALITIES != 0 |
                        shortdata$INJURIES !=0,]

healthdata <- mutate(healthdata, ALLINC = FATALITIES + INJURIES)
healthdata <- select(healthdata, EVTYPE, YEAR, ALLINC)
healthdata <- aggregate(ALLINC*YEAR ~ EVTYPE, healthdata,sum)
names(healthdata) <- c("EVTYPE", "ALLINC")
healthgraph <- ggplot(healthdata, aes(x=EVTYPE, y=ALLINC))+
     geom_bar(stat="identity", fill = "blue")+
     labs(title = "Number of Injuries/Fatalities by Event Type", x= "", y =
"Injuries/Fatalities")+
     theme_bw()+
     theme(axis.text.x=element_text(angle=90))
print(healthgraph)
```



As evidenced in the graph above, TORNADO has been the most harmful weather event causing highest number of injuries and/or fatalities.

Now, let's take a look at the economic data. We will look at the total damage, which we will define as the sum of property and crop damage.

```
econdata <- shortdata[shortdata$PROPDMG !=0|
                          shortdata$CROPDMG !=0,]
econdata <- mutate(econdata, ALLDMG = CROPDMG+PROPDMG)
econdata <- select(econdata, EVTYPE, YEAR, ALLDMG)
econdata <- aggregate(YEAR*ALLDMG~EVTYPE, econdata,sum)
names(econdata) <- c("EVTYPE", "ALLDMG")
econgraph <- ggplot(econdata, aes(x=EVTYPE, y=ALLDMG))+
        geom_bar(stat="identity", fill = "blue")+
        labs(title = "Property and Crop Damage by Event Type",
             x= "", y = "Property and Crop Loss Amount")+
        theme_bw()+
        theme(axis.text.x=element_text(angle=90))
print(econgraph)
```



From the graph above, we can see that WIND and FLOOD are the most harmful economic events causing maximum amount of economic loss.