Reproducible Research Course Project 2

Tikam Singh

20 November 2016

STUDY REPORT - Public Health and Economic Problems Due Severe Weather Events in US

1 - Synopsis

This is a study report of U.S. National Oceanic and Atmospheric Administration's (NOAA) database. The aim of this study is to find out the worst weather event type in terms of population health and economic consequences in US from 1950 to 2011. The analysis shows Tornado was the most harmful weather event type in terms of population health while Flood was the weather event type had the worst economic impact.

2 - Data Processing

2.1 - Loading Data

- The data for this study come from the U.S. National Oceanic and Atmospheric Administration's (NOAA) database.The data can be obtained from the course web site: * Strom Data.
- The events in the database start in the year 1950 and end in November 2011.
- The definition & construction of variables are available from National Weather Service Storm Data Documentation.
- This study is conducted in R Studio (Version 0.99.491). The required packages are ggplot2:

```
knitr::opts_chunk$set(echo = TRUE)
if(!require(ggplot2)) install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```
if(!require(dplyr)) installed.packages("dplyr")
```

```
## Loading required package: 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
```

```
if(!require(readr)) installed.packages("readr")
```

```
## Loading required package: readr
```

```
    * Download data from source and load data into dataframe:
```

```
data <-  "repdata data StormData.csv.bz2"
if (!file.exists(data)) {

        download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",data)
}
if(!exists('storm_data')) {
        storm_data <- read.csv("repdata data StormData.csv.bz2", stringsAsFactors = FALSE)
}
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec =
## dec, : EOF within quoted string
```

2.2 - Subseting Data

```
    * Explore and have a brief idea of the dataset:
```

```
str(storm_data)
```

```
## 'data.frame':    422982 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : chr  "4/18/1950 0:00:00" "4/18/1950 0:00:00" "2/20/1951 0:00:00" "6/8/1951 0:00:00" .
##  $ BGN_TIME  : chr  "0130" "0145" "1600" "0900" ...
##  $ TIME_ZONE : chr  "CST" "CST" "CST" "CST" ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: chr  "MOBILE" "BALDWIN" "FAYETTE" "MADISON" ...
##  $ STATE     : chr  "AL" "AL" "AL" "AL" ...
##  $ EVTYPE    : chr  "TORNADO" "TORNADO" "TORNADO" "TORNADO" ...
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ BGN_AZI   : chr  "" "" "" "" ...
##  $ BGN_LOCATI: chr  "" "" "" "" ...
##  $ END_DATE  : chr  "" "" "" "" ...
##  $ END_TIME  : chr  "" "" "" "" ...
##  $ COUNTY_END: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ COUNTYENDN: logi  NA NA NA NA NA NA ...
##  $ END_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ END_AZI   : chr  "" "" "" "" ...
##  $ END_LOCATI: chr  "" "" "" "" ...
##  $ LENGTH    : num  14 2 0.1 0 0 1.5 1.5 0 3.3 2.3 ...
##  $ WIDTH     : num  100 150 123 100 150 177 33 33 100 100 ...
##  $ F         : int  3 2 2 2 2 2 2 1 3 3 ...
##  $ MAG       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ FATALITIES: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ INJURIES  : num  15 0 2 2 2 6 1 0 14 0 ...
##  $ PROPDMG   : num  25 2.5 25 2.5 2.5 2.5 2.5 2.5 25 25 ...
##  $ PROPDMGEXP: chr  "K" "K" "K" "K" ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: chr  "" "" "" "" ...
##  $ WFO       : chr  "" "" "" "" ...
##  $ STATEOFFIC: chr  "" "" "" "" ...
##  $ ZONENAMES : chr  "" "" "" "" ...
##  $ LATITUDE  : num  3040 3042 3340 3458 3412 ...
##  $ LONGITUDE : num  8812 8755 8742 8626 8642 ...
##  $ LATITUDE_E: num  3051 0 0 0 0 ...
##  $ LONGITUDE_: num  8806 0 0 0 0 ...
##  $ REMARKS   : chr  "" "" "" "" ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

```
            * I subset only the desire variables for the following study, which are the event types ("E
```

```r
subset <- subset(storm_data ,select =
                         c("EVTYPE","FATALITIES","INJURIES","PROPDMG","PROPDMGEXP","CROPDMG", "CROPDMGE
                 EVTYPE != "?" & (FATALITIES >  0 | INJURIES > 0 | PROPDMG > 0 |
```

2.3 - Cleaning Event Type Data

```
    * Check out the construction of Event Type:
```

```r
head(unique(subset$EVTYPE), 20)
```

```
##  [1] "TORNADO"                      "TSTM WIND"
##  [3] "HAIL"                         "ICE STORM/FLASH FLOOD"
##  [5] "WINTER STORM"                 "HURRICANE OPAL/HIGH WINDS"
##  [7] "THUNDERSTORM WINDS"           "HURRICANE ERIN"
##  [9] "HURRICANE OPAL"               "HEAVY RAIN"
## [11] "LIGHTNING"                    "THUNDERSTORM WIND"
## [13] "DENSE FOG"                    "RIP CURRENT"
## [15] "THUNDERSTORM WINS"            "FLASH FLOODING"
## [17] "FLASH FLOOD"                  "TORNADO F0"
## [19] "THUNDERSTORM WINDS LIGHTNING" "THUNDERSTORM WINDS/HAIL"
```

```r
length(unique(subset$EVTYPE))
```

```
## [1] 457
```

```r
subset$EVTYPE <- toupper(subset$EVTYPE)
  # Then I combine similar event types.
  subset$EVTYPE <- gsub('.*HEAT.*', 'HEAT', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*WARM.*', 'HEAT', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*HIGH.*TEMP.*', 'EXTREME HEAT', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*RECORD HIGH TEMPERATURES.*', 'EXTREME HEAT', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*STORM.*', 'STORM', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*FLOOD.*', 'FLOOD', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*WIND.*', 'WIND', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*TORNADO.*', 'TORNADO', subset$EVTYPE)
  subset$EVTYPE <- gsub('.*CLOUD.*', 'CLOUD', subset$EVTYPE)
```

```
subset$EVTYPE <- gsub('.*MICROBURST.*', 'MICROBURST', subset$EVTYPE)
subset$EVTYPE <- gsub('.*BLIZZARD.*', 'BLIZZARD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*COLD.*', 'COLD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*SNOW.*', 'COLD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*FREEZ.*', 'COLD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*LOW TEMPERATURE RECORD.*', 'COLD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*ICE.*', 'COLD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*FROST.*', 'COLD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*LO.*TEMP.*', 'COLD', subset$EVTYPE)
subset$EVTYPE <- gsub('.*HAIL.*', 'HAIL', subset$EVTYPE)
subset$EVTYPE <- gsub('.*DRY.*', 'DRY', subset$EVTYPE)
subset$EVTYPE <- gsub('.*DUST.*', 'DUST', subset$EVTYPE)
subset$EVTYPE <- gsub('.*RAIN.*', 'RAIN', subset$EVTYPE)
subset$EVTYPE <- gsub('.*LIGHTNING.*', 'LIGHTNING', subset$EVTYPE)
subset$EVTYPE <- gsub('.*SUMMARY.*', 'SUMMARY', subset$EVTYPE)
subset$EVTYPE <- gsub('.*WET.*', 'WET', subset$EVTYPE)
subset$EVTYPE <- gsub('.*FIRE.*', 'FIRE', subset$EVTYPE)
subset$EVTYPE <- gsub('.*FOG.*', 'FOG', subset$EVTYPE)
subset$EVTYPE <- gsub('.*VOLCANIC.*', 'VOLCANIC', subset$EVTYPE)
subset$EVTYPE <- gsub('.*SURF.*', 'SURF', subset$EVTYPE)
```

```
length(unique(subset$EVTYPE))
```

```
## [1] 87
```

2.4 - Cleaning Economic Data

```
    * Check out the construction of "PROPDMGEXP" & "CROPDMGEXP":
```

```
 table(subset$PROPDMGEXP)
```

```
## 
##            -      +      0      2      3      4      5      6      7      B      h 
##   7729      1      5    210      1      1      4     18      3      3     12      1 
##      H      K      m      M 
##      6  98818      7   6149 
```

```
table(subset$CROPDMGEXP)
```

```
## 
##            ?      0      B      k      K      m      M 
##  99983      6     17      5     21  12019      1    916 
```

```
    * I convert variable "PROPDMGEXP" & "CROPDMGEXP" into multiplier factors and calculate the cost of
```

```
# Convert all strings to upper case:
subset$PROPDMGEXP <- toupper(subset$PROPDMGEXP)
subset$CROPDMGEXP <- toupper(subset$CROPDMGEXP)

# Update the multiplier factor of PROPDMG
```

```
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='')|(subset$PROPDMGEXP=='-')|(subset$PROPDMGEXP=='?')|(subset$
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='1')] <- 1
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='2')] <- 2
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='3')] <- 3
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='4')] <- 4
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='5')] <- 5
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='6')] <- 6
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='7')] <- 7
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='8')] <- 8
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='H')] <- 2
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='K')] <- 3
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='M')] <- 6
subset$PROPDMGEXP1[(subset$PROPDMGEXP=='B')] <- 9

# Update the multiplier factor of CROPDMG
subset$CROPDMGEXP1[(subset$CROPDMGEXP=='')|(subset$CROPDMGEXP=='-')|(subset$CROPDMGEXP=='?')|(subset$
subset$CROPDMGEXP1[(subset$CROPDMGEXP=='K')] <- 3
subset$CROPDMGEXP1[(subset$CROPDMGEXP=='M')] <- 6
subset$CROPDMGEXP1[(subset$CROPDMGEXP=='B')] <- 9

# Calculate cost of damages
subset$PROPDMGCOST <- subset$PROPDMG*10^as.numeric(subset$PROPDMGEXP1)
subset$CROPDMGCOST <- subset$CROPDMG*10^as.numeric(subset$CROPDMGEXP1)
```

2.5 - Summarizing Data

```
    * Create a new dataset to summarize the Health Impact (total number of fatalities and injures), the
```

```
subset2 <- aggregate( x = list(Health_Impact = subset$FATALITIES + subset$INJURIES),
                      by=list(EVENT_TYPE=subset$EVTYPE),
                      FUN=sum, na.rm=TRUE)
subset2 <- subset2[order(subset2$Health_Impact, decreasing=T),]
```

```
    * Create a new dataset to summarize the Damage Cost (total cost of property damage and crop damage)
```

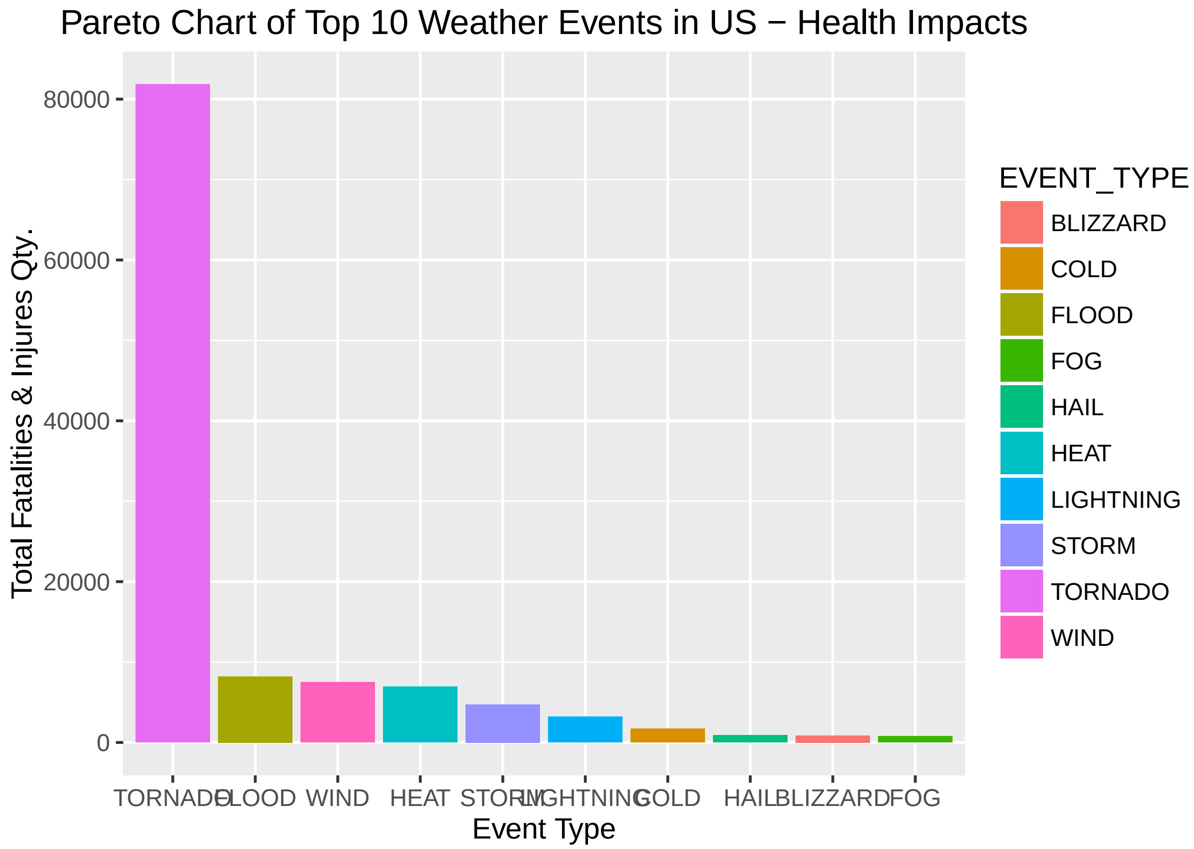
```
subset3 <- aggregate( x = list(Damage_Cost = subset$PROPDMGCOST + subset$CROPDMGCOST),
                      by=list(EVENT_TYPE=subset$EVTYPE),
                      FUN=sum, na.rm=TRUE)
subset3 <- subset3[order(subset3$Damage_Cost, decreasing=T),]
```

3.1 - Find out the top 10 harmful weather event type with respect to population health

```
    * Produce the pareto chart of the worst 10 weather events in terms of population     health:
```

```
HealthImpactChart <- ggplot(head(subset2,10), aes(x=reorder(EVENT_TYPE, -Health_Impact), y=Health_Impa
                       geom_bar(stat="identity") +
                       xlab("Event Type") + ylab("Total Fatalities & Injures Qty.") +
                       ggtitle("Pareto Chart of Top 10 Weather Events in US - Health Impacts")
```

```r
print(HealthImpactChart)
```



*Show the exact number of health impacts of above chart:

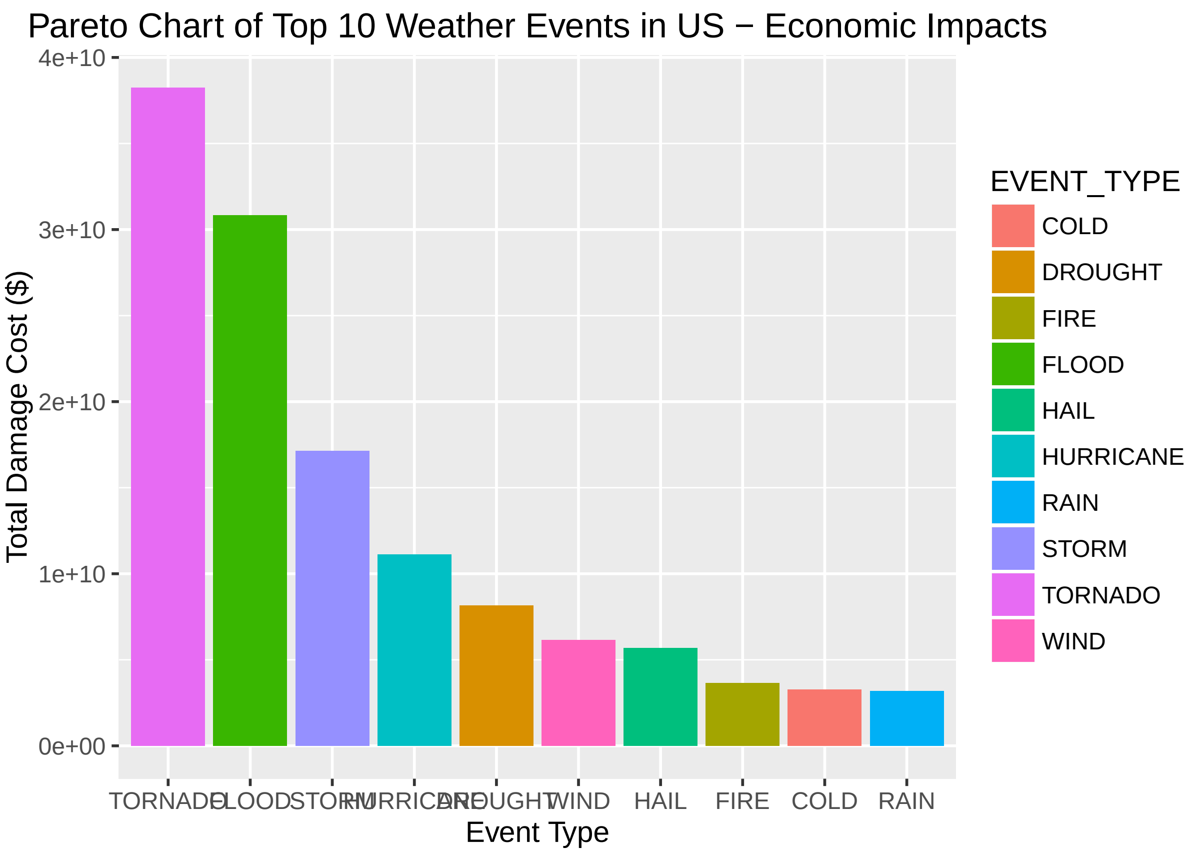
```r
head(subset2,10)
```

```
##    EVENT_TYPE Health_Impact
## 73    TORNADO         81824
## 15      FLOOD          8218
## 84       WIND          7485
## 20       HEAT          6938
## 71      STORM          4731
## 49  LIGHTNING          3193
## 9        COLD          1715
## 19       HAIL           903
## 5    BLIZZARD           872
## 16        FOG           789
```

3.2 - Find out the top 10 weather event types have the greatest economic consequences

```
*pareto chart of the worst 10 weather events in terms of economic consequences:
```

```r
DamageCostChart <- ggplot(head(subset3,10), aes(x=reorder(EVENT_TYPE, -Damage_Cost), y=Damage_Cost, f
                      geom_bar(stat="identity") +
                      xlab("Event Type") + ylab("Total Damage Cost ($)") +
                    ggtitle("Pareto Chart of Top 10 Weather Events in US - Economic Impacts")
print(DamageCostChart)
```



```
    * The exact damage cost of above chart:
```

```
head(subset3,10)
```

```
##    EVENT_TYPE Damage_Cost
## 73    TORNADO 38236131076
## 15      FLOOD 30833444235
## 71      STORM 17146568364
## 32  HURRICANE 11123814000
## 12    DROUGHT  8154079000
## 84       WIND  6136858438
## 19       HAIL  5692337656
## 14       FIRE  3637219700
## 9        COLD  3255664820
## 61       RAIN  3195399950
```

4 Conclusions

The outcome of NOAA database study between 1950 and 2011 shows the worst weather events type are presented as below: ## 4.1 - Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

```
* Tornado was the most harmful weather event type in terms of population health.
```

4.2 - Across the United States, which types of events have the greatest economic consequences?

```
* Flood was the worst weather event type in terms of economic consequences.
```