Impact of Severe Weather on Public Health and Economy in the United States

Alexander Kuznetsov 5/3/2018

Synopsis

The purpose of this study is to analyze impact of severe weather based on data collected by NOAA between 1950 and 2011 in the United States. Top ten devastating weather events are identified by their effect on public health and economy. Health impact is estimated by calculating number of injuries and fatalities for each type of severe weather event. Property and crop damage are used as gauges for economic impact from adverse weather. Although, the goal of the project is to analyze existing data without making any changes, closer look at the dataset allows us to make major revision on the damage caused by severe weather events.

Data Processing

NOAA dataset is to be loaded to the working directory. After working directory in R is set, file can be read using read.csv function into data frame "df" with spaces used for separation and keeping original header titles. Functions such as dim, head, tail, summary and str can be very instrumental to look at the dataset.

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 3.4.4
```

```
opts_chunk$set(tidy.opts=list(width.cutoff=65),tidy=TRUE)
```

```
setwd("C:/Users/Ulpan/Documents/Coursera/DataScience/Notes/Reproducible Research/Project 2")
df <- read.csv("repdata%2Fdata%2FStormData.csv.bz2", sep = ",", header = TRUE)
dim(df)
```

```
## [1] 902297     37
```

In order to estimate weather impact, following columns can be selected:

```
names(df)[c(2, 3, 7, 8, 22, 23, 24, 25, 26, 27, 28)]
```

```
##  [1] "BGN_DATE"   "BGN_TIME"   "STATE"      "EVTYPE"     "MAG"
##  [6] "FATALITIES" "INJURIES"   "PROPDMG"    "PROPDMGEXP" "CROPDMG"
## [11] "CROPDMGEXP"
```

Data frame with selected columns "sdf" will be used throughout this project.

```
sdf <- df[, c(2, 3, 7, 8, 22, 23, 24, 25, 26, 27, 28)]
```

Health Impact

Following code is to select only non-zero values for injuries and fatalities from the data frame, store them in varialbe health1. Next, we add up injuries and fatalities for each weather event using aggregate function and arrange the results in descending order.

```
library(dplyr)
```

```
health1 <- sdf[which(sdf$FATALITIES != 0 | sdf$INJURIES != 0), ]
dim(health1)
```

```
## [1] 21929    11
```

```r
health1i <- aggregate(INJURIES ~ EVTYPE, data = health1, sum)
health1i1 <- arrange(health1i, desc(INJURIES))
health1f <- aggregate(FATALITIES ~ EVTYPE, data = health1, sum)
health1f1 <- arrange(health1f, desc(FATALITIES))
```

Let's combine resulting data frames and rename columns which would help when data are plotted and analyzed in the next section:

```r
health2 <- cbind(health1i1, health1f1$FATALITIES)
colnames(health2) <- c("Events", "Injuries", "Fatalities")
head(health2)
```

```
##          Events Injuries Fatalities
## 1       TORNADO    91346       5633
## 2     TSTM WIND     6957       1903
## 3         FLOOD     6789        978
## 4 EXCESSIVE HEAT     6525        937
## 5     LIGHTNING     5230        816
## 6          HEAT     2100        504
```

These are how top 10 categories which cause most injuries and fatalities look like:

```r
health3 <- health2[1:10, ]
health3
```

```
##              Events Injuries Fatalities
## 1           TORNADO    91346       5633
## 2         TSTM WIND     6957       1903
## 3             FLOOD     6789        978
## 4    EXCESSIVE HEAT     6525        937
## 5         LIGHTNING     5230        816
## 6              HEAT     2100        504
## 7         ICE STORM     1975        470
## 8       FLASH FLOOD     1777        368
## 9  THUNDERSTORM WIND     1488        248
## 10             HAIL     1361        224
```

Economic Impact

Summaries of PROPDMGEXP and CROPDMGEXP columns show breakdown of letter notations for damage cost.

```r
summary(sdf$PROPDMGEXP)
```

```
##             -      ?      +      0      1      2      3      4      5
## 465934      1      8      5    216     25     13      4      4     28
##      6      7      8      B      h      H      K      m      M
##      4      5      1     40      1      6 424665      7  11330
```

```r
summary(sdf$CROPDMGEXP)
```

```
##             ?      0      2      B      k      K      m      M
## 618413      7     19      1      9     21 281832      1   1994
```

Following code converts letter notations into numbers in PROPDMG and CROPDMG. The result is stored in variable cost1:

```r
cost1 <- sdf
cost1$PROPDMG <- ifelse(cost1$PROPDMGEXP == "h", cost1$PROPDMG * 100,
    cost1$PROPDMG)
cost1$PROPDMG <- ifelse(cost1$PROPDMGEXP == "H", cost1$PROPDMG * 100,
    cost1$PROPDMG)
cost1$PROPDMG <- ifelse(cost1$PROPDMGEXP == "K", cost1$PROPDMG * 1000,
    cost1$PROPDMG)
cost1$PROPDMG <- ifelse(cost1$PROPDMGEXP == "M", cost1$PROPDMG * 1e+06,
    cost1$PROPDMG)
cost1$PROPDMG <- ifelse(cost1$PROPDMGEXP == "m", cost1$PROPDMG * 1e+06,
    cost1$PROPDMG)
cost1$PROPDMG <- ifelse(cost1$PROPDMGEXP == "B", cost1$PROPDMG * 1e+09,
    cost1$PROPDMG)
cost1$CROPDMG <- ifelse(cost1$CROPDMGEXP == "K", cost1$CROPDMG * 1000,
    cost1$CROPDMG)
cost1$CROPDMG <- ifelse(cost1$CROPDMGEXP == "k", cost1$CROPDMG * 1000,
    cost1$CROPDMG)
cost1$CROPDMG <- ifelse(cost1$CROPDMGEXP == "m", cost1$CROPDMG * 1e+06,
    cost1$CROPDMG)
cost1$CROPDMG <- ifelse(cost1$CROPDMGEXP == "M", cost1$CROPDMG * 1e+06,
    cost1$CROPDMG)
cost1$CROPDMG <- ifelse(cost1$CROPDMGEXP == "B", cost1$CROPDMG * 1e+09,
    cost1$CROPDMG)
```

Number notations in PROPDMGEXP and CROPDMGEXP comprise insignificant number of records and will be ignored. Next, check for the changes in PROPDMG and CROPDMG:

```r
summary(cost1$PROPDMG)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 0.000e+00 0.000e+00 4.736e+05 5.000e+02 1.150e+11
```

```r
summary(cost1$CROPDMG)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## 0.000e+00 0.000e+00 0.000e+00 5.442e+04 0.000e+00 5.000e+09
```

Next, let's filter out zero values in PROPDMG and CROPDMG, create new column - TOTALDMG capturing total economic damage by adding property damage and crop damage.

```r
cost2 <- filter(cost1, PROPDMG != 0 | CROPDMG != 0)
dim(cost2)
```

```
## [1] 245031     11
```

```r
cost2$TOTALDMG <- cost2$PROPDMG + cost2$CROPDMG
```

Selecting non-zero values helped reduce number of rows by more than 3 times. Now, data frame stored in cost2 can be aggregated by total damage, property damage and crop damage using sum function. Processed data are to be stored in the following variables: cost3, cost2PDagr and cost2CDagr respectively.

```r
cost3 <- aggregate(TOTALDMG ~ EVTYPE, data = cost2, sum)
cost2PDagr <- aggregate(PROPDMG ~ EVTYPE, data = cost2, sum)
cost2CDagr <- aggregate(CROPDMG ~ EVTYPE, data = cost2, sum)
head(cost3)
```

```
##                   EVTYPE TOTALDMG
## 1    HIGH SURF ADVISORY   200000
```

```
## 2           FLASH FLOOD    50000
## 3             TSTM WIND  8100000
## 4       TSTM WIND (G45)     8000
## 5                     ?     5000
## 6   AGRICULTURAL FREEZE 28820000
```

```r
head(cost2PDagr)
```

```
##                  EVTYPE PROPDMG
## 1    HIGH SURF ADVISORY  200000
## 2           FLASH FLOOD   50000
## 3             TSTM WIND 8100000
## 4       TSTM WIND (G45)    8000
## 5                     ?    5000
## 6   AGRICULTURAL FREEZE       0
```

```r
head(cost2CDagr)
```

```
##                  EVTYPE  CROPDMG
## 1    HIGH SURF ADVISORY        0
## 2           FLASH FLOOD        0
## 3             TSTM WIND        0
## 4       TSTM WIND (G45)        0
## 5                     ?        0
## 6   AGRICULTURAL FREEZE 28820000
```

These 3 outputs of aggregate function can now be combined into one data frame - costdf and arranged in descending order. Just for convenience of plotting data in the next section, we will select top 10 rows and rename columns storring data in costdfpl2.

```r
costdf <- cbind(cost2PDagr, cost2CDagr$CROPDMG, cost3$TOTALDMG)
head(costdf)
```

```
##                  EVTYPE PROPDMG cost2CDagr$CROPDMG cost3$TOTALDMG
## 1    HIGH SURF ADVISORY  200000                  0         200000
## 2           FLASH FLOOD   50000                  0          50000
## 3             TSTM WIND 8100000                  0        8100000
## 4       TSTM WIND (G45)    8000                  0           8000
## 5                     ?    5000                  0           5000
## 6   AGRICULTURAL FREEZE       0           28820000       28820000
```

```r
colnames(costdf)[3:4] <- c("CROPDMG", "TOTALDMG")
costdf <- arrange(costdf, desc(TOTALDMG))
costdfpl2 <- costdf[1:10, ]
colnames(costdfpl2) <- c("Event", "Property Damage", "Crop Damage",
    "Total Damage")
costdfpl2
```

```
##                Event Property Damage Crop Damage Total Damage
## 1              FLOOD    144657709807  5661968450 150319678257
## 2  HURRICANE/TYPHOON     69305840000  2607872800  71913712800
## 3            TORNADO     56937160779   414953270  57352114049
## 4        STORM SURGE     43323536000        5000  43323541000
## 5               HAIL     15732267543  3025954473  18758222016
## 6        FLASH FLOOD     16140812067  1421317100  17562129167
## 7            DROUGHT      1046106000 13972566000  15018672000
## 8          HURRICANE     11868319010  2741910000  14610229010
```

```
## 9        RIVER FLOOD      5118945500  5029459000  10148404500
## 10         ICE STORM      3944927860  5022113500   8967041360
```

Results

Health Impact

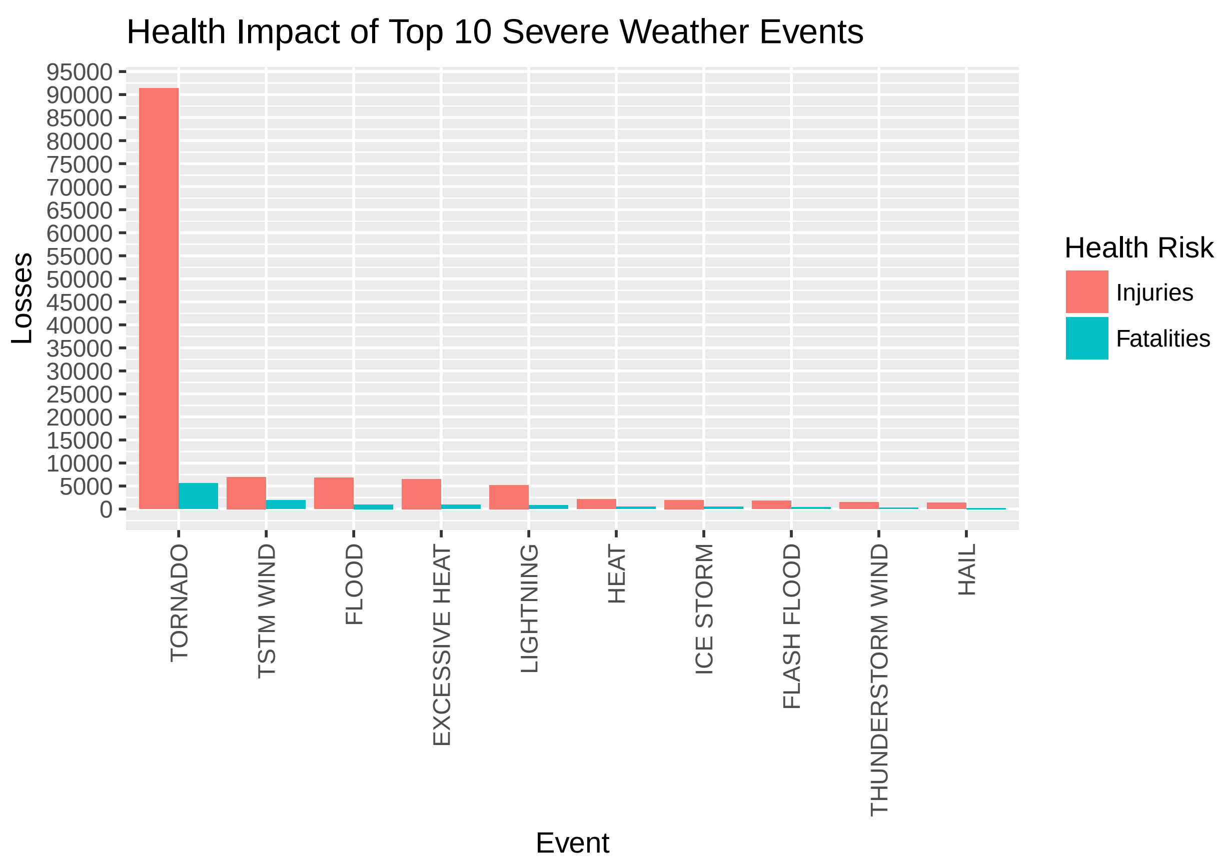
As seen before, tornadoes create most significant danger to public health. Their impact outnumbers any other types of severe weather by order of magnitude. In order to put these results into perspective, we will plot data breaking down impact from each weather event into two risk categories: injuries and fatalities.

```
library(reshape2)
library(ggplot2)
```

```
health3m <- melt(health3)
```

```
## Using Events as id variables
```

```
colnames(health3m)[2:3] <- c("Risk", "Losses")
ggplot(data = health3m, aes(y = Losses, fill = Risk, x = reorder(Events,
    -Losses))) + geom_bar(stat = "identity", position = position_dodge()) +
    theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 0.5)) +
    labs(y = "Losses", fill = "Health Risk", x = "Event") + scale_y_continuous(breaks = seq(0,
    1e+05, 5000)) + ggtitle("Health Impact of Top 10 Severe Weather Events")
```



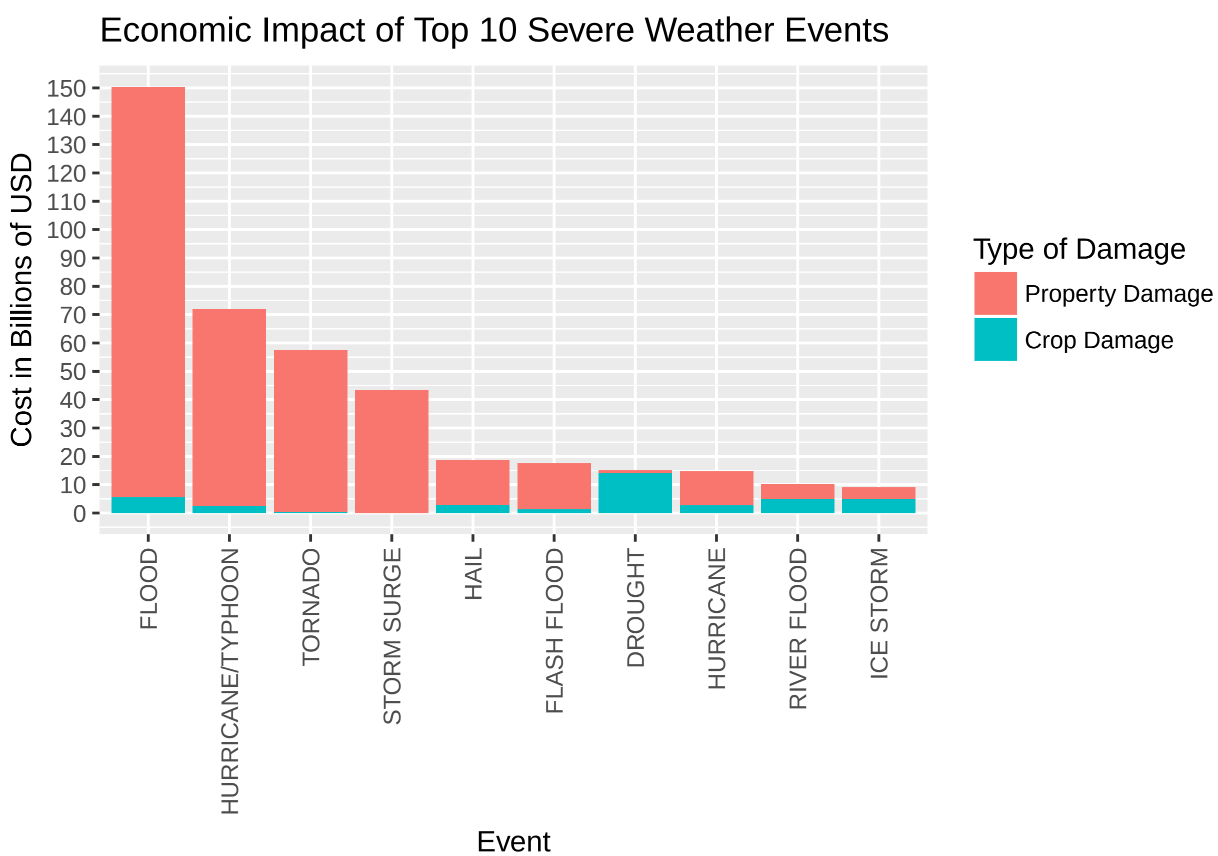
Further improvements to the analysis of health risks can also be made. For example, there are two event categories for winds associated with thunderstorms: “THUNDERSTORM WIND” and “TSTM WIND” which can be combined into one category. Same probably applies to "HEAT" and "EXCESSIVE HEAT". Although such changes would not change the overall picture, there would definetely be quantitative changes ot the outcome. Such 'data cleaning' can be a goal for separate project.

Economic Impact

```
costdfpl3 <- costdfpl2[, 1:3]
costdfpl3m <- melt(costdfpl3)
```

```
## Using Event as id variables
```

```
colnames(costdfpl3m)[2:3] <- c("Damage", "Cost")
costdfpl3m$Cost <- costdfpl3m$Cost/1e+09
ggplot(data = costdfpl3m, aes(y = Cost, fill = Damage, x = reorder(Event,
    -Cost))) + geom_bar(stat = "identity") + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 0.5)) + labs(y = "Cost in Billions of USD",
    fill = "Type of Damage", x = "Event") + scale_y_continuous(breaks = seq(0,
    200, 10)) + ggtitle("Economic Impact of Top 10 Severe Weather Events")
```



Floods, hurricanes, tornadoes are most impactful in terms of property damage as well as overall cost to economy. Most agricultural damage occurred during droughts, river floods and ice storms. Similar to the previous discussion on health impact, some 'cleaning' of data can be done on the NOAA dataset. For example, "HURRICANE/TYPHOON" and "HURRICANE" event categories can be combined into one. These events are the same except for their geographic origin: hurricanes form over Atlantic Ocean and Caribbean Sea while typhoons occur in Pacific Ocean. Next section addresses some of these issues as well as incostistency in data.

Additional Comments

After witnessing both hurricane and major flood in Houston area during last 10 years I was not surprised at all that these events cause most damage. Hurricane Ike which impacted Houston in September 2008 devastated city with strong winds and severe rain. Area I lived in did not have power for almost 3 weeks. After almost 10 years, my only memories of the hurricane are fallen trees and traffic lights, filled up rivers and canals, and painfully loud noise from the very strong wind outside which lasted whole night. The sound of the wind was similar to the noise from jet engine set to full power. When Ike hit Houston it was category 2 or 3 hurricane. I really cannot imagine what happens in area impacted by category 5 hurricane. Devastation must be much more severe. Hurricane Harvey in August 2017 did not have powerful winds, but brought a lot of precipitation into the area. It got stuck for few days between Houston and Gul of Mexico coast, sucking moisture from the Gulf and dumping it to the coastal area rsulting in major flood. The memories now are almost non-stop rain for 4 days, overfilled lakes in the neighbourhood, military helicopters zipping back and forth, staying home for almost a week because office was closed. Our area was not impacted much, while others were not so lucky. Although, impact was huge from this historical flood, and, unfortunately, people lost their lives, it still seems that major hurricane can cause more economic devastation to the area than flood. That made me go through data one more time to look at major floods and hurricanes in NOAA dataset.

$100 Billion Problem

First thing that jumped out was this:

```
cost1[which(cost1$PROPDMG == max(cost1$PROPDMG)), ]
```

```
##                 BGN_DATE    BGN_TIME STATE EVTYPE MAG FATALITIES INJURIES
## 605953 1/1/2006 0:00:00 12:00:00 AM    CA  FLOOD   0          0        0
##         PROPDMG PROPDMGEXP  CROPDMG CROPDMGEXP
## 605953 1.15e+11          B 32500000          M
```

The highest property damage value in entire dataset was a record for flood event in California which amounted to $115 billion. This single event accounts for the majority of the dmage caused by floods between 1950 and 2011 which totals to around 145 billion. For example, damage from hurricane Harvey is estimated to be around $130 billion. However, description in the REMARKS column gives an impression of flood that had much smaller impact. In addition, different estimate for property damage is mentioned for this event:

```
cafld <- df[which(cost1$PROPDMG == max(cost1$PROPDMG)), ]
cafld$REMARKS
```

```
## [1] Major flooding continued into the early hours of January 1st, before the Napa River finally fell
## 436781 Levels:  -2 at Deer Park\n ... Zones 22 and 23 were added to the high wind warning of  January
```

Here is full description of the event:

```
cafld
```

```
##        STATE__         BGN_DATE    BGN_TIME TIME_ZONE COUNTY COUNTYNAME
## 605953       6 1/1/2006 0:00:00 12:00:00 AM       PST     55       NAPA
##        STATE EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI         END_DATE
## 605953    CA  FLOOD         0         COUNTYWIDE 1/1/2006 0:00:00
##           END_TIME COUNTY_END COUNTYENDN END_RANGE END_AZI END_LOCATI
## 605953 07:00:00 AM          0         NA         0         COUNTYWIDE
##        LENGTH WIDTH  F MAG FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG
## 605953      0     0 NA   0          0        0     115          B    32.5
##        CROPDMGEXP WFO          STATEOFFIC ZONENAMES LATITUDE LONGITUDE
## 605953          M MTR CALIFORNIA, Western              3828     12218
##        LATITUDE_E LONGITUDE_
## 605953       3828      12218
```

```
##
## 605953 Major flooding continued into the early hours of January 1st, before the Napa River finally fe
##        REFNUM
## 605953 605943
```

After searching online, it appears that total damage from this flood was estimated at $300 million: https://pubs.usgs.gov/of/2006/1182/pdf/ofr2006-1182.pdf It made me thinking that "B" was entered by mistake instead of "M" in PROPDMGEXP column. This would significantly change our result for economic damage. Code below is used to fix the problem with this input. Same steps as described above in "Data Processing" and "Results" sectoins are followed to come up with final result. Therefore, I will skip text and provide only code.

```r
cost1[605953, 8:9] <- c(1.15e+08, "M")
cost1[605953, ]
```

```
##                BGN_DATE    BGN_TIME STATE EVTYPE MAG FATALITIES INJURIES
## 605953 1/1/2006 0:00:00 12:00:00 AM    CA  FLOOD   0          0        0
##         PROPDMG PROPDMGEXP  CROPDMG CROPDMGEXP
## 605953 1.15e+08          M 32500000          M
```

```r
cost1a <- cost1
cost2a <- cost1a[which(cost1a$PROPDMG != 0 | cost1a$CROPDMG != 0),
    ]
cost2a$PROPDMG <- as.numeric(cost2a$PROPDMG)
cost2a$CROPDMG <- as.numeric(cost2a$CROPDMG)
cost2a$TOTALDMG <- cost2a$PROPDMG + cost2a$CROPDMG
cost3a <- aggregate(TOTALDMG ~ EVTYPE, data = cost2a, sum)
cost3aPDagr <- aggregate(PROPDMG ~ EVTYPE, data = cost2a, sum)
cost3aCDagr <- aggregate(CROPDMG ~ EVTYPE, data = cost2a, sum)
cost4a <- cbind(cost3aPDagr, cost3aCDagr$CROPDMG, cost3a$TOTALDMG)
colnames(cost4a) <- c("Event", "PROPDMG", "CROPDMG", "TOTALDMG")
cost5a <- arrange(cost4a, desc(TOTALDMG))
head(cost5a)
```

```
##               Event     PROPDMG    CROPDMG    TOTALDMG
## 1 HURRICANE/TYPHOON 69305840000 2607872800 71913712800
## 2           TORNADO 56937160779  414953270 57352114049
## 3       STORM SURGE 43323536000       5000 43323541000
## 4             FLOOD 29772709807 5661968450 35434678257
## 5              HAIL 15732267543 3025954473 18758222016
## 6       FLASH FLOOD 16140812067 1421317100 17562129167
```

```r
colnames(cost5a)[2:4] <- c("Property Damage", "Crop Damage", "Total Damage")
costdfa <- cost5a[1:10, ]
costdfa
```

```
##                 Event Property Damage Crop Damage Total Damage
## 1  HURRICANE/TYPHOON     69305840000  2607872800  71913712800
## 2            TORNADO     56937160779   414953270  57352114049
## 3        STORM SURGE     43323536000        5000  43323541000
## 4              FLOOD     29772709807  5661968450  35434678257
## 5               HAIL     15732267543  3025954473  18758222016
## 6        FLASH FLOOD     16140812067  1421317100  17562129167
## 7            DROUGHT      1046106000 13972566000  15018672000
## 8          HURRICANE     11868319010  2741910000  14610229010
## 9        RIVER FLOOD      5118945500  5029459000  10148404500
## 10         ICE STORM      3944927860  5022113500   8967041360
```

Let's now combine damage from hurricanes and typhoons:

```
cost4a$Event <- as.character(cost4a$Event)
grep("HURRICANE|Hurricane|hurricane", cost4a$Event, value = TRUE)
```

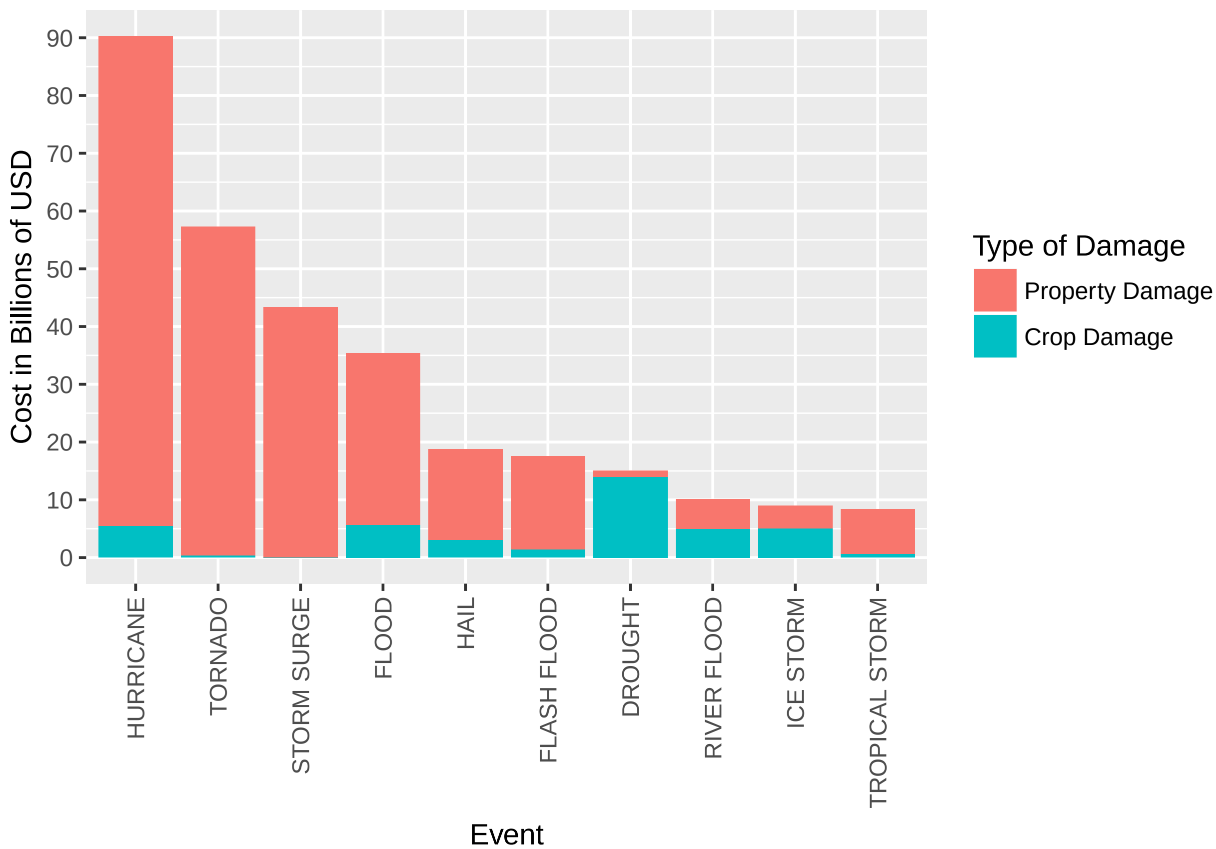
```
## [1] "HURRICANE"                  "HURRICANE-GENERATED SWELLS"
## [3] "HURRICANE EMILY"            "HURRICANE ERIN"
## [5] "HURRICANE FELIX"            "HURRICANE GORDON"
## [7] "HURRICANE OPAL"             "HURRICANE OPAL/HIGH WINDS"
## [9] "HURRICANE/TYPHOON"
```

Besides categories for hurricane and typhoon, there are few individual hurricanes that were storred as separate categories, such as Hurricane Emily or Erin. These categories are consolidated under one category - "HURRICANES" below, but similar cases can be found for many other categories. Consolidation for all of these cases could be done as separate project.

```
costcon <- cost4a
costcon$Event <- ifelse(grepl("HURRICANE", cost4a$Event) == TRUE,
    "HURRICANE", cost4a$Event)
costconagr <- aggregate(TOTALDMG ~ Event, data = costcon, sum)
costconPDagr <- aggregate(PROPDMG ~ Event, data = costcon, sum)
costconCDagr <- aggregate(CROPDMG ~ Event, data = costcon, sum)
costcondf <- cbind(costconPDagr, costconCDagr$CROPDMG, costconagr$TOTALDMG)
colnames(costcondf)[3:4] <- c("CROPDMG", "TOTALDMG")
costcondf <- arrange(costcondf, desc(costcondf$TOTALDMG))
costcondf1 <- costcondf
colnames(costcondf1)[2:3] <- c("Property Damage", "Crop Damage")
costcondf2 <- costcondf1[1:10, 1:3]
costcondf2m <- melt(costcondf2)
```

```
## Using Event as id variables
```

```
colnames(costcondf2m)[2:3] <- c("Damage", "Cost")
costcondf2m$Cost <- costcondf2m$Cost/1e+09
ggplot(data = costcondf2m, aes(y = Cost, fill = Damage, x = reorder(Event,
    -Cost))) + geom_bar(stat = "identity") + theme(axis.text.x = element_text(angle = 90,
    hjust = 1, vjust = 0.5)) + labs(y = "Cost in Billions of USD",
    fill = "Type of Damage", x = "Event") + scale_y_continuous(breaks = seq(0,
    200, 10))
```



Conclusion

The most dangerous events for human life and health are found to be tornadoes. At the same time, hurricanes/typhoons and floods cause most of property damage. Drought is the source of most impact on agriculture. It could be shown that NOAA dataset contains error related to the property damage for the flood in Napa Valley of California in December 2005 - January 2006. Revised results for the property damage are presented in this report. It appears that most damage is caused by hurricanes and tornadoes followed by storm surges and floods.