Peer Assignment 2 of the Course Reproducible Research- Part of Coursera's John Hopkins University Data Science Specialization Series

by Stefan Schmager

Sunday, April 26, 2015

Assignment Goal

The basic goal of this assignment is to explore the U.S. National Oceanic and Atmospheric Administration's (NOAA) Storm Database and answer some basic questions about severe weather events. The database must be used to answer a pair of research questions and show the code for the entire analysis. The analysis can consist of tables, figures, or other summaries. Any R package may be used to support the analysis.

The Most Harmful Weather Events in the United States from 1950 to 2011 in Terms of Damage to Population Health and Economic Consequences

Synopsis

This analysis is based on publicly available data from the U.S. National Oceanic and Atmospheric Administration's which is downloaded and read into R, the utilized statistical programming software. The original dataset is stripped from irrelevant variables and limited to those variables of interest for the underlying research questions about the most harmful weather events in the US in terms of population health and economic consequences. Dataset records with events outside the country were filtered out. Further, damage measures to determine the most harmful weather events were computed. Last, one graph and table answer each research question and summarize the top ten most harmful weather events according to their computed damage measure. A few remarks describe the limitations of the analysis and give ideas for further analyses.

Introduction

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, property, and crop damage, and preventing such outcomes to the extent possible is a key concern.

This project involves exploring the NOAA storm database. This database tracks characteristics of major storms and weather events in the United States, including when and where they occur, as well as estimates of any fatalities, injuries, and economic damage.

Data

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. The file is downloaded from the linked web site below:

- Storm Data [47Mb]

There is also some documentation of the database available. Here, one may find how some of the variables are constructed and defined.

- National Weather Service Storm Data Documentation
- National Climatic Data Center Storm Events FAQ
- Additional Code Book from IRE (Investigative Reporters & Editors)

The events in the database start in the year 1950 and end in November 2011. In the earlier years of the database there are generally fewer events recorded, most likely due to a lack of good records. More recent years should be considered more complete.

Data Processing

The following packages are loaded and used in the analysis. The session information are also provided were versions and loaded packages are summarized.

```
# Load packages
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(tidyr)
library(datasets)
library(ggplot2)
library(xlsx)
```

```
## Loading required package: rJava

## Loading required package: xlsxjars
```

```
sessionInfo()
```

```
## R version 3.2.3 (2015-12-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## locale:
```

```
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] xlsx_0.5.7     xlsxjars_0.6.1 rJava_0.9-8    ggplot2_2.0.0
## [5] tidyr_0.4.1    dplyr_0.4.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.3      knitr_1.12.3     magrittr_1.5     munsell_0.4.2
##  [5] colorspace_1.2-6 R6_2.1.1         stringr_1.0.0    plyr_1.8.3
##  [9] tools_3.2.3      parallel_3.2.3   grid_3.2.3       gtable_0.1.2
## [13] DBI_0.3.1        htmltools_0.3    yaml_2.1.13      assertthat_0.1
## [17] digest_0.6.9     formatR_1.2.1    evaluate_0.8     rmarkdown_0.9.5
## [21] stringi_1.0-1    scales_0.3.0
```

The data set is downloaded from the indicated data source, which is locally saved as a CSV data file in the designated working directory and read into R for further processing.

```
DataSource  <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
DataFile    <- "StormData.csv"
DataPlace   <- "~/GitHub/OnlineCourses/Coursera/Data-Science Specialization by Johns Hopkins University/
setwd(DataPlace)

if (!file.exists(DataFile)) {
      download.file(url = "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2",
                    destfile = DataFile)
      }; DownloadDate = date()

DownloadDate
```

```
## [1] "Wed Apr 27 05:47:11 2016"
```

```
dat <- read.csv(file = DataFile)
```

The relevant variables for answering the research questions are selected.

```
dat <- select(dat,
              STATE, EVTYPE, BGN_DATE,    # U.S. state, event type and date
              FATALITIES, INJURIES,       # measures for harm of population
              PROPDMG, PROPDMGEXP,        # measures for (economic) proporty damage
              CROPDMG, CROPDMGEXP         # measures for (economic) crop damage
              )
```

The event-date variable is transformed from a character to a date variable, and summarized by its distribution measures. The summary of the date variable shows how the database records are distributed over time. For instance, half of the database records were entered after the displayed **median** date. The database records that serve as inputs of the analysis span from the **min.** to the **max.** date summarized below.

```
dat$DATE <- as.Date(dat$BGN_DATE, "%m/%d/%Y")
summary(dat$DATE)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "1950-01-03" "1995-04-20" "2002-03-18" "1998-12-27" "2007-07-28"
##         Max.
## "2011-11-30"
```

The data set contains records with state abbreviations that go beyond the Unites States. Therefore, the records must be filtered and limited to those entries that escribe one of the 50 state abbreviations.

```
# Display all states in the initial dataset
unique(dat$STATE)
```

```
##  [1] AL AZ AR CA CO CT DE DC FL GA HI ID IL IN IA KS KY LA ME MD MA MI MN
## [24] MS MO MT NE NV NH NJ NM NY NC ND OH OK OR PA RI SC SD TN TX UT VT VA
## [47] WA WV WI WY PR AK ST AS GU MH VI AM LC PH GM PZ AN LH LM LE LS SL LO
## [70] PM PK XX
## 72 Levels: AK AL AM AN AR AS AZ CA CO CT DC DE FL GA GM GU HI IA ID ... XX
```

```
# Compare to vector with 50 U.S. states from the R "state" data set
data(state)
factor(unique(state.abb))
```

```
##  [1] AL AK AZ AR CA CO CT DE FL GA HI ID IL IN IA KS KY LA ME MD MA MI MN
## [24] MS MO MT NE NV NH NJ NM NY NC ND OH OK OR PA RI SC SD TN TX UT VT VA
## [47] WA WV WI WY
## 50 Levels: AK AL AR AZ CA CO CT DE FL GA HI IA ID IL IN KS KY LA MA ... WY
```

```
# Filter out the 50 U.S. states from the dataset
dat <- dat[dat$STATE %in% state.abb,]
dat$STATE <- factor(dat$STATE)

# Compare all states in the filtered dataset to the same result at beginning of code chunk
unique(dat$STATE)
```

```
##  [1] AL AZ AR CA CO CT DE FL GA HI ID IL IN IA KS KY LA ME MD MA MI MN MS
## [24] MO MT NE NV NH NJ NM NY NC ND OH OK OR PA RI SC SD TN TX UT VT VA WA
## [47] WV WI WY AK
## 50 Levels: AK AL AR AZ CA CO CT DE FL GA HI IA ID IL IN KS KY LA MA ... WY
```

The measures for damage in terms of **population health** are relatively simply defined by **counts of directly killed (fatalities) and directly injured human beings (injuries)**.

However, the measures for **economic damage** are defined and computed more complicated according to the database documentation on page 12, _“[..] Damage estimates should be entered as actual dollar amounts [...] Estimates should be rounded to three significant digits, followed by an alphabetical character signifying the magnitude of the number, i.e., 1.55B for $1,550,000,000. Alphabetical characters used to signify magnitude include”K" for thousands, “M” for millions, and “B” for billions.“_

```
## Measures in terms of population health are defined

# FATALITIES (Number of directly killed)
dat$FATALITIES <- as.integer(dat$FATALITIES)

# INJURIES (Number of directly injured)
dat$INJURIES <- as.integer(dat$INJURIES)

## Measures in terms of economic damage are defined and compute

# PROPDMG/EXP (Property damage in whole numbers and hundredths & A multiplier where Hundred (H), Thousa
dat$PROPDMGEXP <- factor(toupper(dat$PROPDMGEXP))

summary(dat$PROPDMGEXP)
```

```
##             -      ?      +      0      1      2      3      4      5
## 456190      1      8      5    216     25     13      4      4     28
##      6      7      8      B      H      K      M
##      4      5      1     39      7 415375  11261
```

```
dat$PROPDMGEXP2[dat$PROPDMGEXP == ""]  <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "-"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "?"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "+"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "0"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "1"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "2"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "3"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "4"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "5"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "6"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "7"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "8"] <- 1
dat$PROPDMGEXP2[dat$PROPDMGEXP == "H"] <- 100
dat$PROPDMGEXP2[dat$PROPDMGEXP == "K"] <- 1000
dat$PROPDMGEXP2[dat$PROPDMGEXP == "M"] <- 1000000
dat$PROPDMGEXP2[dat$PROPDMGEXP == "B"] <- 1000000000

# Computed measure
dat$PropertyDamage <- dat$PROPDMG*dat$PROPDMGEXP2

# CROPDMG/EXP (Crop damage in whole numbers and hundredths & A multiplier where Hundred (H), Thousand (
dat$CROPDMGEXP <- factor(toupper(dat$CROPDMGEXP))

summary(dat$CROPDMGEXP)
```

```
##             ?      0      2      B      K      M
## 607897      7     19      1      9 273283   1970
```

```
dat$CROPDMGEXP2[dat$CROPDMGEXP == ""]  <- 1
dat$CROPDMGEXP2[dat$CROPDMGEXP == "?"] <- 1
dat$CROPDMGEXP2[dat$CROPDMGEXP == "0"] <- 1
```

```
dat$CROPDMGEXP2[dat$CROPDMGEXP == "2"] <- 1
dat$CROPDMGEXP2[dat$CROPDMGEXP == "H"] <- 100
dat$CROPDMGEXP2[dat$CROPDMGEXP == "K"] <- 1000
dat$CROPDMGEXP2[dat$CROPDMGEXP == "M"] <- 1000000
dat$CROPDMGEXP2[dat$CROPDMGEXP == "B"] <- 1000000000

# Computed measure
dat$CropDamage <- dat$CROPDMG*dat$CROPDMGEXP2
```

Note: For full correctness of the analysis, the recorded event types should have been coded to more uniform and distinct descriptions. The displayed event-type records provide an example for the messy data entries: **"thunderstorm winds"** and **"tstm wind"** describe the same event, namely **"Thunderstorm (Wind)"**; however, due to non-uniform descriptions the entries cannot be aggregated as such. The information from the documentation on pages 5 and 6 may be helpful in categorizing the event types, *"The only events permitted in Storm Data are listed in Table 1 of Section 2.1.1. The chosen event name should be the one that most accurately describes the meteorological event leading to fatalities, injuries, damage, etc."*

```
dat$EVTYPE <- factor(toupper(dat$EVTYPE))
head(unique(dat$EVTYPE), 10)
```

```
##  [1] TORNADO                   TSTM WIND
##  [3] HAIL                      FREEZING RAIN
##  [5] SNOW                      ICE STORM/FLASH FLOOD
##  [7] SNOW/ICE                  WINTER STORM
##  [9] HURRICANE OPAL/HIGH WINDS THUNDERSTORM WINDS
## 867 Levels:  COASTAL FLOOD  FLASH FLOOD  LIGHTNING ... WND
```

The examples listed below show how the recorded event types may be categorized by more uniform event-type descriptions to fascilitate further data aggregation.

```
# dat$EventType[dat$EVTYPE == ' coastal flood']   <- 'Coastal Flood'
# dat$EventType[dat$EVTYPE == ' flash flood']     <- 'Flash Flood'
# dat$EventType[dat$EVTYPE == ' lightning']       <- 'Lightning'
# dat$EventType[dat$EVTYPE == ' tstm wind']       <- 'Thunderstorm (Wind)'
# dat$EventType[dat$EVTYPE == ' tstm wind (g45)'] <- 'Thunderstorm (Wind)'
# dat$EventType[dat$EVTYPE == ' waterspout']      <- 'Waterspout'
# dat$EventType[dat$EVTYPE == '?']                <- 'Others'
```

The dataset, where records with states beyond the U.S. had been filtered out already, is now aggregated per recorded event type across the 50 states and across time from the earliest recorded date and the latest recorded date.

```
dat0 <- as.data.frame(summarize(group_by(dat, EVTYPE),
                                CropDamage     = sum(CropDamage,     na.rm = T),
                                PropertyDamage = sum(PropertyDamage, na.rm = T),
                                Injuries       = sum(INJURIES,       na.rm = T),
                                Fatalities     = sum(FATALITIES,     na.rm = T)))
```

The main measures for **economic** and **population-health** damage are computed by **adding crop and property damage (both measured in USD)**, as well as **fatalities and injuries (both measured in counts of human beings injured/killed)** respectively.

The main measures for economic and population-health damage are also truncated (divided by a million and a thousand respetcively) so that high numbers do not inflate tables or graphs later.

Measures for **economic damage** such as crop and property damage, as well as the sum of both are **measured in billion US dollars**.

Measures for **population-health damage** such as injuries and fatalities, as well as the sum of both are **measured in counts of thousand**.

```
# Economic Damage
dat0$EconDamage <- dat0$CropDamage + dat0$PropertyDamage
summary(dat0$EconDamage)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
##            0            0            0    544500000       100000
##         Max.
## 150100000000
```

```
summary(dat0$CropDamage)
```

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##           0           0           0    55800000           0 13970000000
```

```
summary(dat0$PropertyDamage)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
##            0            0            0    488700000        75000
##         Max.
## 144500000000
```

```
dat0$CropDamage          <- dat0$CropDamage/    1000000000 # in billions
dat0$PropertyDamage      <- dat0$PropertyDamage/1000000000 # in billions
dat0$EconDamage          <- dat0$EconDamage/    1000000000 # in billions

# Population-Health Damage
dat0$PopDamage           <- dat0$Fatalities + dat0$Injuries
summary(dat0$PopDamage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0     178       0   96980
```

```
summary(dat0$Fatalities)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   17.11    0.00 5633.00
```

```
summary(dat0$Injuries)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0     0.0   160.8     0.0 91350.0
```

```
dat0$PopDamage           <- dat0$PopDamage/ 1000 # in thousands
dat0$Fatalities          <- dat0$Fatalities/1000 # in thousands
dat0$Injuries            <- dat0$Injuries/  1000 # in thousands
```

Data Analysis

The data set is split in two separate ones in order to further answer both research questions.

The event types in each set are also ordered by their total damage measure and filtered to the top 10 most harmful event types. Further, the data sets need to be reshaped (measure columns are gathered and condensed into a column pair of measures and values) in order to be visualized subsequently.

```
dat1        <- select(dat0, EVTYPE, Fatalities, Injuries, PopDamage) %>%
               arrange(desc(PopDamage)) %>% head(10)

table1      <- dat1

dat1        <- gather(select(dat1, -PopDamage), "Measure", "Value", Fatalities, Injuries)

graph1      <- ggplot(dat1, aes(x = reorder(EVTYPE, Value), y = Value, fill = Measure)) +
               geom_bar(stat = "identity", position = "stack") +
               theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
               coord_flip() +
               ggtitle("Event Types with the Greatest Population Health Damage") +
               xlab("Event Types") +
               ylab("Measure Counts (in Thousands)") +
               theme(legend.position = "bottom")
```

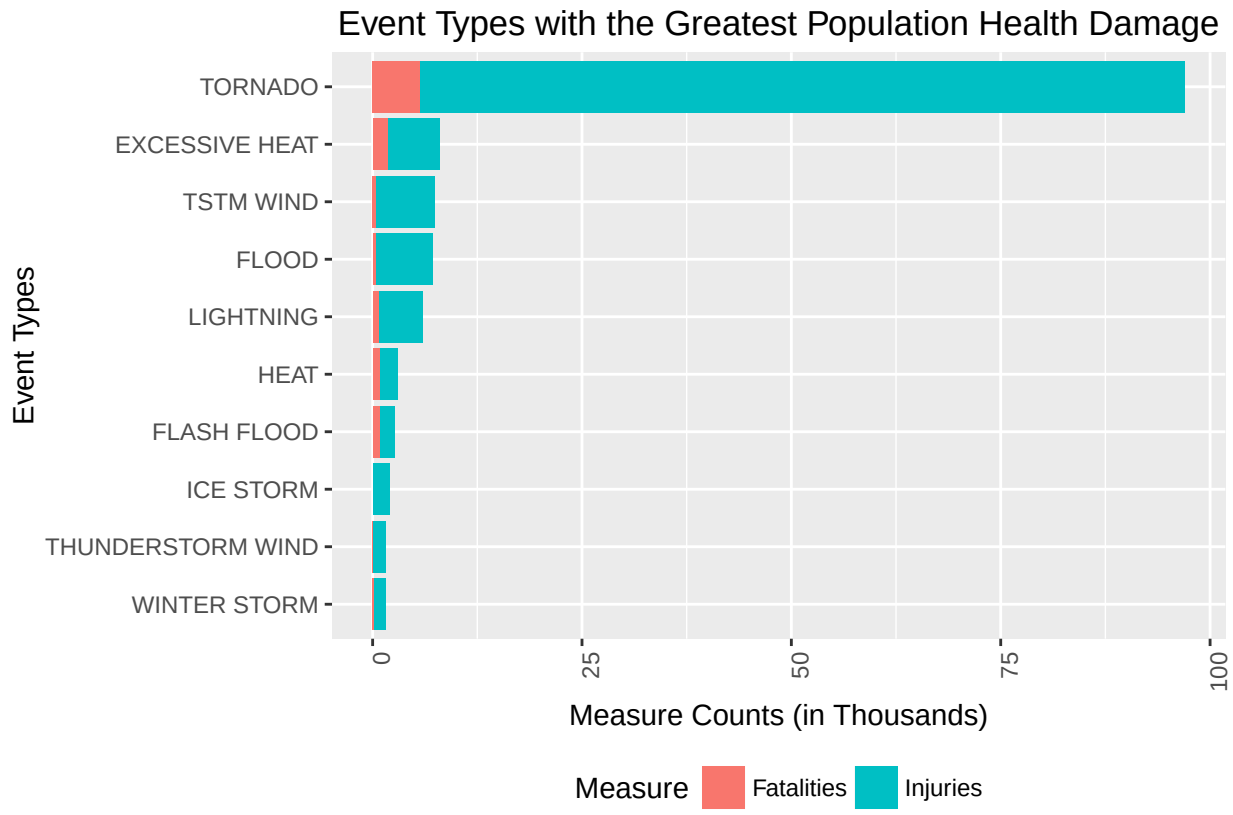
```
dat2        <- select(dat0, EVTYPE, EconDamage, PropertyDamage, CropDamage) %>%
               arrange(desc(EconDamage)) %>% head(10)

table2      <- dat2

dat2        <- gather(select(dat2, -EconDamage), "Measure", "Value", PropertyDamage, CropDamage)

graph2      <- ggplot(dat2,
               aes(x = reorder(EVTYPE, Value), y = Value, fill = Measure)) +
               geom_bar(stat = "identity", position = "stack") +
               theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
               coord_flip() +
               ggtitle("Event Types with the Greatest Economic Consequences") +
               xlab("Event Types") +
               ylab("Measure (in Billion USD)") +
               theme(legend.position = "bottom")
```

Results

1. Across the United States, which (top ten) event types are most harmful with respect to population health?

As one can see in the graph and table below, most human beings get harmed by **tornados**. That is mostly due to a high count in people injured and less so due to individuals killed.

```
graph1
```



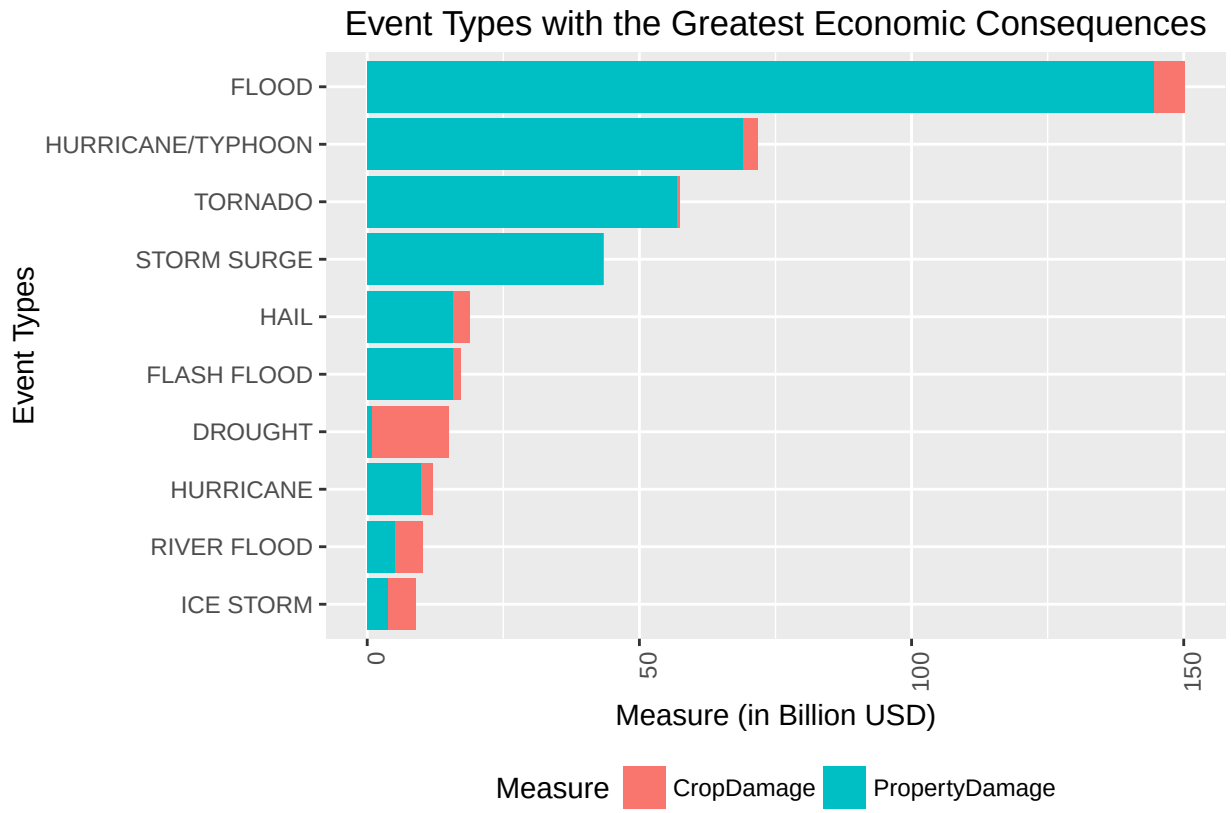

```
rename(table1, WeatherEvent = EVTYPE, TotalDamage = PopDamage)
```

```
##          WeatherEvent Fatalities Injuries TotalDamage
## 1             TORNADO      5.633   91.346      96.979
## 2      EXCESSIVE HEAT      1.883    6.209       8.092
## 3           TSTM WIND      0.502    6.941       7.443
## 4               FLOOD      0.464    6.786       7.250
## 5           LIGHTNING      0.806    5.212       6.018
## 6                HEAT      0.935    2.100       3.035
## 7         FLASH FLOOD      0.939    1.767       2.706
## 8           ICE STORM      0.089    1.975       2.064
## 9   THUNDERSTORM WIND      0.133    1.471       1.604
## 10       WINTER STORM      0.205    1.321       1.526
```

2. Across the United States, which (top ten) event types have the greatest economic consequences?

As one can see in the graph and table below, most economic damage is caused by **floods**. That is mostly due to a high damage in property and less so due to crop damage.

```
graph2
```




```
rename(table2, WeatherEvent = EVTYPE, TotalDamage = EconDamage)
```

```
##          WeatherEvent TotalDamage PropertyDamage CropDamage
## 1               FLOOD  150.145287     144.531319  5.6139684
## 2  HURRICANE/TYPHOON   71.636601      69.033100  2.6035008
## 3             TORNADO   57.351641      56.936689  0.4149523
## 4         STORM SURGE   43.323466      43.323461  0.0000050
## 5                HAIL   18.758215      15.732261  3.0259545
## 6         FLASH FLOOD   17.275076      15.868170  1.4069051
## 7             DROUGHT   15.013467       1.041106 13.9723610
## 8           HURRICANE   12.103928       9.913998  2.1899300
## 9         RIVER FLOOD   10.148401       5.118942  5.0294590
## 10          ICE STORM    8.967021       3.944908  5.0221135
```

Remarks For Future Analyses

Per assignment instructions, the types of weather events are indicated by the variable EVTYPE. It contains the individual descriptions of the ones who entered the weather event into the database. Further research should encompass a correct categorization of those event descriptions into more uniform categories. As one can see, the abbreviation TSTM WIND and THUNDERSTORM WIND describe the same weather event, but are listed individually.

Also, further research should pay tribute to the skewed event date distribution. Early years of database entries may want to be filtered out to allow more precision for recent data entries.

Last, it may be interesting to see how certain regions or particular states of the country are affected by weather events and how they compare.