Title: “Reproducible research: assignment #2”

Author: “Stijn” Start date: “Sunday, February 7, 2016” Output: html_document

Synopsis of Study Results

This project involves exploring the U.S. National Oceanic and Atmospheric Administration’s (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 property damage.

Questions this study considers

1. Across the United States, which types of events (EVTYPE variable) are most harmful with respect to population health?
2. Across the United States, which types of events have the greatest economic consequences?

Notes about the compute environment that was used

This study was done using the following tools, including OS and Programming language versions

MACHINE: 64-bit; Windows 7 Pro SP1 machine with 4 cores; 8GB RAM.

SOFTWARE: R language: RStudio Version 0.98.1091 – © 2009-2014 RStudio, Inc.

```
Github reference for this project: https://github.com/Stijnvaneven/RepData_PeerAssignment2
```

Set libraries used in this analysis

```
library(stringr)
library(data.table)
library(dplyr)
library(ggplot2)
library(reshape2)
library(gridExtra)
```

Loading the 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. You can download the file from the course web site:

Storm Data [47Mb] There is also some documentation of the database available. Here you will

find how some of the variables are constructed/defined.

National Weather Service Storm Data Documentation National Climatic Data Center Storm Events FAQ 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.

```
StormData_Url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormDat
a.csv.bz2"
StormData_Zip <- "data/StormData.csv.bz2"
StormData_Rds <- "data/StormData.RDS"

if (!file.exists(StormData_Zip)) {
    download.file(url = StormData_Url,
                  destfile = StormData_Zip)
}

## For faster processing, check for R Data Set save file for subsequent runs of s
cript.
RDSloaded <- FALSE
if (!file.exists(StormData_Rds)) {
    SD <- read.csv(file = bzfile(StormData_Zip), strip.white = TRUE)
    # save data to uncompressed csv file.
    # write.csv(SD, file = "data/StormData.csv")
    saveRDS(SD, file = "data/StormData.RDS")
} else {
    SD <- readRDS(StormData_Rds)
    RDSloaded <- TRUE
}
```

Data Processing

The following variables (see line of code) are of interest to our study. I am creating a smaller data frame with just those columns to speed up computations.

```
DSsubset<-subset(SD, select = c("EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "P
ROPDMGEXP", "CROPDMG", "CROPDMGEXP"))
```

How Many weather event types are there?

```
summarize(DSsubset, n_distinct(EVTYPE))
```

```
##   n_distinct(EVTYPE)
## 1                985
```

Question 1: Find the total number

of fatalities and injuries by event type

```
totalFatalities <- aggregate(FATALITIES~EVTYPE,DSsubset,sum)
totalInjuries <- aggregate(INJURIES~EVTYPE,DSsubset,sum)

## Combine the 2 data frames
InjuriesFatalitiesDF<-merge(totalFatalities,totalInjuries)

## order the dataframe by number of fatalities. There are 935 type of events.
## Pick only the top 10 with highest number of fatalities
InjuriesFatalitiesDF10 <- data.table(InjuriesFatalitiesDF[order(InjuriesFatalitie
sDF$FATALITIES, decreasing = TRUE), ][1:10, ])

## insert an index column as the first column and order the colums
InjuriesFatalitiesDF10$index <- c(1:nrow(InjuriesFatalitiesDF10))
setcolorder(InjuriesFatalitiesDF10, c("index", "EVTYPE", "INJURIES", "FATALITIE
S"))
```

Results for question 1

print and plot

```
## print the entire table.
print("IMPACT ON INJURIES AND FATALITIES BY EVENT - TOP 10")
```

```
## [1] "IMPACT ON INJURIES AND FATALITIES BY EVENT - TOP 10"
```

```
print(InjuriesFatalitiesDF10, row.names = FALSE)
```

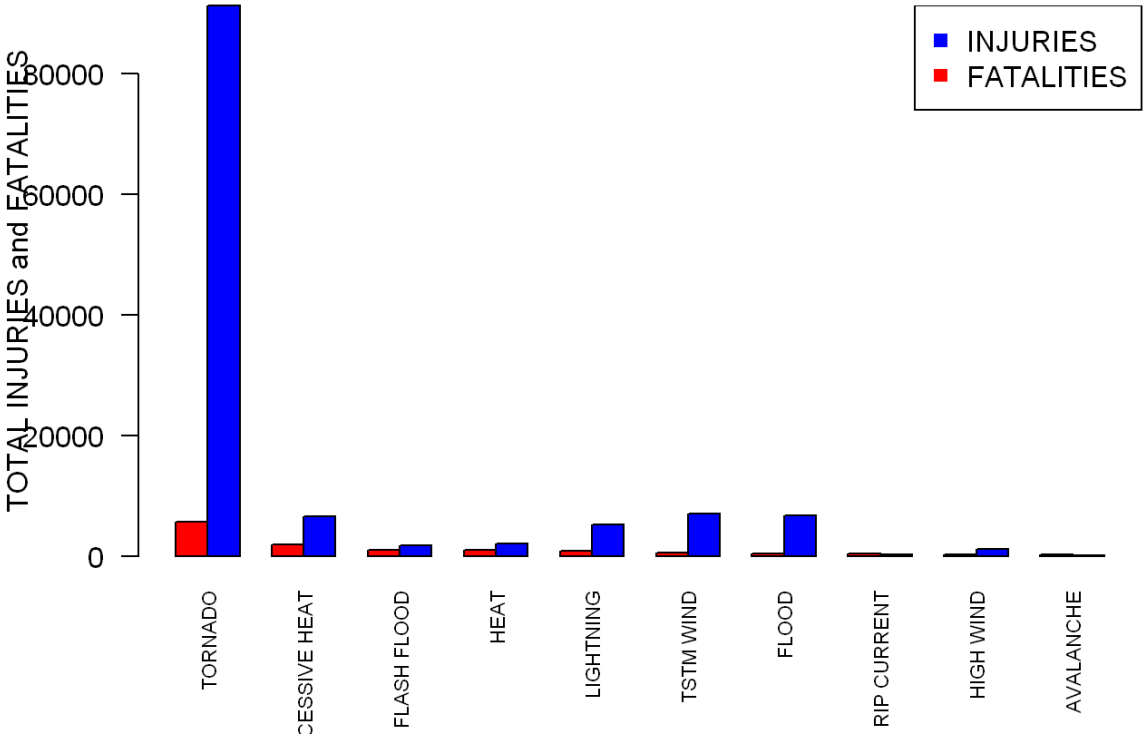
```
##  index         EVTYPE INJURIES FATALITIES
##      1        TORNADO    91346       5633
##      2 EXCESSIVE HEAT     6525       1903
##      3    FLASH FLOOD     1777        978
##      4           HEAT     2100        937
##      5      LIGHTNING     5230        816
##      6      TSTM WIND     6957        504
##      7          FLOOD     6789        470
##      8    RIP CURRENT      232        368
##      9      HIGH WIND     1137        248
##     10      AVALANCHE      170        224
```

```
## We will use barplots to display the results of the table. Display
```

```
## both the injuries and fatalities on the same plot
x <- rbind(InjuriesFatalitiesDF10$FATALITIES, InjuriesFatalitiesDF10$INJURIES)

barplot(x, beside = TRUE, las = 2, cex.names= 0.7, col = c("red", "blue"), ylim =
c(0, max(InjuriesFatalitiesDF10$INJURIES)), names.arg = InjuriesFatalitiesDF10$EVT
YPE, ylab = "TOTAL INJURIES and FATALITIES")

legend("topright", c("INJURIES", "FATALITIES"), col = c("blue", "red"), pch = 15)
```



Question 2: Inspect which weather events generate the most economic damage.

```
##Select the rows that have a "billion" dollar PROP damage multiple
BillionsPropertyList <- DSsubset[,"PROPDMGEXP"] == "B"
BillionsPropertySubset <- DSsubset[BillionsPropertyList,]

## Check top 10 property expenses
TopBilProp <- top_n(BillionsPropertySubset,10,PROPDMG)
TopBilProp
```

```
##               EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG
## 1       WINTER STORM          4        0    5.00          B    0.00
```

```
## 2        RIVER FLOOD          0        0    5.00          B    5.00
## 3     TROPICAL STORM         22        0    5.15          B    0.00
## 4  HURRICANE/TYPHOON          7      780    5.42          B  285.00
## 5  HURRICANE/TYPHOON          5        0   10.00          B    0.00
## 6  HURRICANE/TYPHOON          0        0   16.93          B    0.00
## 7        STORM SURGE          0        0   31.30          B    0.00
## 8  HURRICANE/TYPHOON          0        0    7.35          B    0.00
## 9        STORM SURGE          0        0   11.26          B    0.00
## 10 HURRICANE/TYPHOON         15      104    5.88          B    1.51
## 11             FLOOD          0        0  115.00          B   32.50
##    CROPDMGEXP
## 1
## 2           B
## 3
## 4           M
## 5
## 6
## 7
## 8
## 9
## 10          B
## 11          M
```

```
##Select the rows that have a "billion" dollar CROP damage multiple
BillionsCropList <- DSsubset[,"CROPDMGEXP"] == "B"
BillionsCropSubset <- DSsubset[BillionsCropList,]

## Check top 10 crop expenses
TopBilCrop <- top_n(BillionsCropSubset,10,CROPDMG)
TopBilCrop
```

```
##              EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG
## 1              HEAT          0        0    0.00               0.40
## 2       RIVER FLOOD          0        0    5.00          B    5.00
## 3           DROUGHT          0        0    0.00               0.50
## 4            FREEZE          0        0    0.00               0.20
## 5         ICE STORM          0        0  500.00          K    5.00
## 6 HURRICANE/TYPHOON         15      104    5.88          B    1.51
## 7           DROUGHT          0        0    0.00               1.00
## 8           DROUGHT          0        0    0.00          K    0.00
## 9           DROUGHT          0        0    0.00          K    0.00
##   CROPDMGEXP
## 1          B
## 2          B
## 3          B
## 4          B
## 5          B
```

```
## 6          B
## 7          B
## 8          B
## 9          B
```

Data Processing to compute the damages

```
## convert the exponent letter symbols into the power digit to use.

#coerce values as characters
DSsubset$PROPDMGEXP <- as.character(DSsubset$PROPDMGEXP)
DSsubset$CROPDMGEXP <- as.character(DSsubset$CROPDMGEXP)

DSsubset[DSsubset$PROPDMGEXP %in% c("+", "-", ""),]$PROPDMGEXP <- 0
DSsubset[DSsubset$CROPDMGEXP %in% c("?", ""),]$CROPDMGEXP <- 0

DSsubset[DSsubset$PROPDMGEXP == "H",]$PROPDMGEXP <- 2

DSsubset[DSsubset$PROPDMGEXP == "K",]$PROPDMGEXP <- 3
DSsubset[DSsubset$CROPDMGEXP == "K",]$CROPDMGEXP <- 3

DSsubset[DSsubset$PROPDMGEXP == "M",]$PROPDMGEXP <- 6
DSsubset[DSsubset$CROPDMGEXP == "M",]$CROPDMGEXP <- 6

DSsubset[DSsubset$PROPDMGEXP == "B",]$PROPDMGEXP <- 9
DSsubset[DSsubset$CROPDMGEXP == "B",]$CROPDMGEXP <- 9

DSsubset$PROPDMGEXP[is.na(DSsubset$PROPDMGEXP)] = 0
DSsubset$CROPDMGEXP[is.na(DSsubset$CROPDMGEXP)] = 0

#coerce values as numeric
DSsubset$PROPDMGEXP<- as.numeric(DSsubset$PROPDMGEXP)
```

```
## Warning: NAs introduced by coercion
```

```
DSsubset$CROPDMGEXP<- as.numeric(DSsubset$CROPDMGEXP)
```

```
## Warning: NAs introduced by coercion
```

```
# compute the damages: apply the power to the damage columns and store into two NE
W columns
DSsubset$realPROPDMG<- DSsubset$PROPDMG*10^DSsubset$PROPDMGEXP
DSsubset$realCROPDMG<- DSsubset$CROPDMG*10^DSsubset$CROPDMGEXP
```

Results for question 2

```
# sum the property costs per event type and sort in descending order
propertyDMG <- aggregate(realPROPDMG~EVTYPE, data=DSsubset, sum)
propertyDMG_desc<- propertyDMG[order(-propertyDMG$realPROPDMG),]
# Subset to top 10 for display
PropertyDMG10<-propertyDMG_desc[1:10,]
PropertyDMG10
```

```
##                EVTYPE  realPROPDMG
## 170             FLOOD 144657709807
## 411 HURRICANE/TYPHOON  69305840000
## 834           TORNADO  56935880677
## 670       STORM SURGE  43323536000
## 153       FLASH FLOOD  16822673979
## 244              HAIL  15730367513
## 402         HURRICANE  11868319010
## 848    TROPICAL STORM   7703890550
## 972      WINTER STORM   6688497251
## 359         HIGH WIND   5270046295
```

```
# sum the crop costs per event type and sort in descending order
cropDMG <- aggregate(realCROPDMG~EVTYPE, data=DSsubset, sum)
cropDMG_desc<- cropDMG[order(-cropDMG$realCROPDMG),]
# Subset to top 10 for display
cropDMG10<-cropDMG_desc[1:10,]
cropDMG10
```

```
##                EVTYPE realCROPDMG
## 95            DROUGHT 13972566000
## 170             FLOOD  5661968450
## 590       RIVER FLOOD  5029459000
## 427         ICE STORM  5022113500
## 244              HAIL  3025537473
## 402         HURRICANE  2741910000
## 411 HURRICANE/TYPHOON  2607872800
## 153       FLASH FLOOD  1421317100
## 140      EXTREME COLD  1292973000
## 212      FROST/FREEZE  1094086000
```

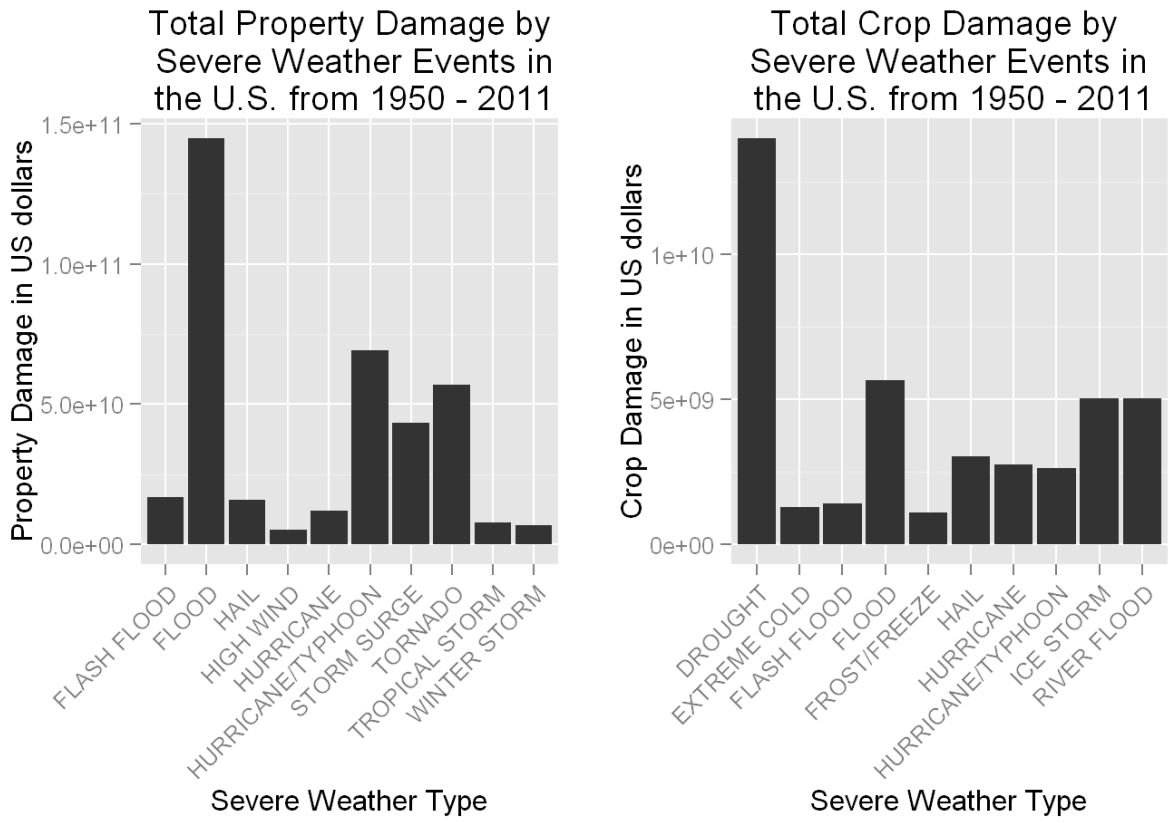
Generate the plots based on the top 10 most costly events

```
propertyPlot <- qplot(EVTYPE, data = PropertyDMG10, weight = realPROPDMG, geom
= "bar", binwidth = 1) + theme(axis.text.x = element_text(angle = 45, hjust =
1)) + scale_y_continuous("Property Damage in US dollars") +
xlab("Severe Weather Type") + ggtitle("Total Property Damage by\n Severe Weather
Events in\n the U.S. from 1950 - 2011")

cropPlot<- qplot(EVTYPE, data = cropDMG10, weight = realCROPDMG, geom = "bar", bi
nwidth = 1) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) + scale_y_continuous
("Crop Damage in US dollars") +
    xlab("Severe Weather Type") + ggtitle("Total Crop Damage by \nSevere Weather
Events in\n the U.S. from 1950 - 2011")
```

Display

```
grid.arrange(propertyPlot, cropPlot, ncol = 2)
```



Conclusion

Answer to question 1: - Tornadoes and droughts are the main causes of deaths. - Tornadoes cause the most injuries by far, followed by floods

Answer to question 2: - Floods and droughts are the main causes of crop damage - Floods and storms are the main causes of property damage.