Course Project 2 - Stormdata

Tom Garshol

12/26/2018

This document presents the data processing, analysis and result of the U.S National Oceanic and Atmospheric Administration???s storm database interms of its impact on property damage and human health.

This project consists of a databaseanalyse , for to find the answeres for two major question:

Which types of events are most harmful with respect to population health, all over USA?

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

The data source -> https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2 (https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2)

The .bz file has been expaned, and repdata-data-StormData.csv will be the source.

Initial coding :

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------- tidyv
erse 1.2.1 --
```

```
## <U+221A> ggplot2 2.2.1     <U+221A> purrr   0.2.4
## <U+221A> tibble  1.4.2     <U+221A> dplyr   0.7.4
## <U+221A> tidyr   0.8.0     <U+221A> stringr 1.2.0
## <U+221A> readr   1.1.1     <U+221A> forcats 0.2.0
```

```
## -- Conflicts -------------------------------------------------- tidyverse_c
onflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
StormData <- read.csv("repdata-data-StormData.csv", sep = ",", stringsAsFactors =
F)
str(StormData)
```

```
## 'data.frame':    902297 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 ...
```

Finding the variables that will be used :

```
df <- StormData[, c("BGN_DATE", "EVTYPE", "FATALITIES", "INJURIES", "PROPDMG", "PR
OPDMGEXP", "CROPDMG", "CROPDMGEXP")]
```

“evtype” will be regrouped for all categories that hasn??t been collected.

```
df$EVENT <- "OTHERS"
df$EVENT[grep(".*HAIL.*", df$EVTYPE, ignore.case = TRUE)] <- "HAIL"
df$EVENT[grep(".*HEAT.*", df$EVTYPE, ignore.case = TRUE)] <- "HEAT"
df$EVENT[grep(".*FIRE.*", df$EVTYPE, ignore.case = TRUE)] <- "HEAT"
df$EVENT[grep(".*HIGH TEMPERATURE.*", df$EVTYPE, ignore.case = TRUE)] <- "HEAT"
df$EVENT[grep(".*COLD.*", df$EVTYPE, ignore.case = TRUE)] <- "COLD"
df$EVENT[grep(".*FROST.*", df$EVTYPE, ignore.case = TRUE)] <- "COLD"
df$EVENT[grep(".*LOW TEMPERATURE.*", df$EVTYPE, ignore.case = TRUE)] <- "COLD"
df$EVENT[grep(".*FLOOD.*", df$EVTYPE, ignore.case = TRUE)] <- "FLOOD"
df$EVENT[grep(".*WIND.*", df$EVTYPE, ignore.case = TRUE)] <- "WIND"
df$EVENT[grep(".*STORM.*", df$EVTYPE, ignore.case = TRUE)] <- "STORM"
df$EVENT[grep(".*SNOW.*", df$EVTYPE, ignore.case = TRUE)] <- "SNOW"
df$EVENT[grep(".*TORNADO.*", df$EVTYPE, ignore.case = TRUE)] <- "TORNADO"
df$EVENT[grep(".*HURRICANE.*", df$EVTYPE, ignore.case = TRUE)] <- "HURRICANE"
df$EVENT[grep(".*LIGHTNING.*", df$EVTYPE, ignore.case = TRUE)] <- "LIGHTNING"
df$EVENT[grep(".*DROUGHT.*", df$EVTYPE, ignore.case = TRUE)] <- "DROUGHT"
df$EVENT[grep(".*FOG.*", df$EVTYPE, ignore.case = TRUE)] <- "FOG"
df$EVENT[grep(".*RAIN.*", df$EVTYPE, ignore.case = TRUE)] <- "RAIN"
```

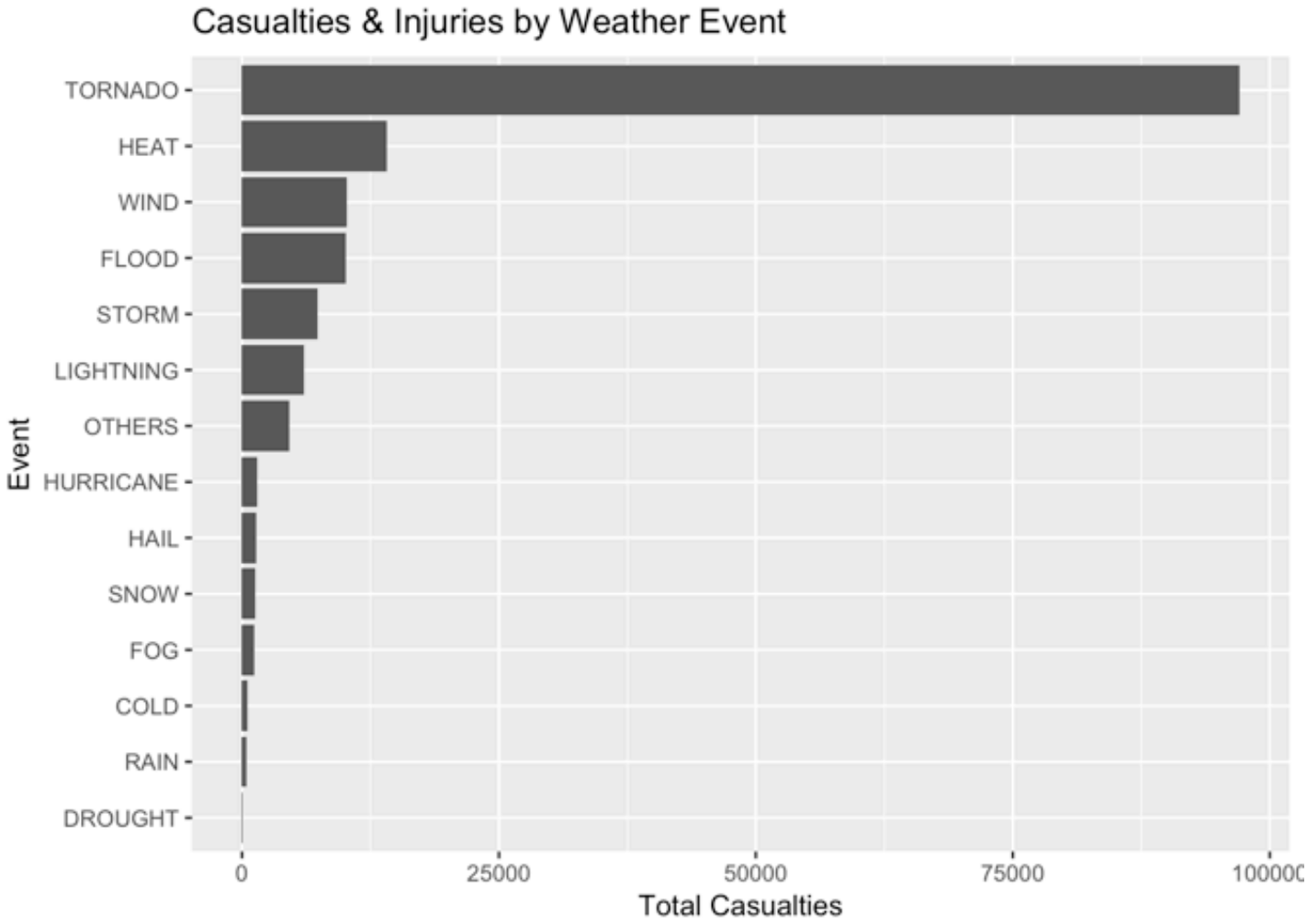
Mapping the property damage to numeric values :

```
key <- sort(unique(df$PROPDMGEXP))
key_value <- c(0,0,0,1,10,10,10,10,10,10,10,10,10,10^9,10^2,10^2,10^3,10^6,10^6)
df_key_value <- data.frame(key, key_value)

df$property_factor <- df_key_value$key_value[match(df$PROPDMGEXP, df_key_value$key
)]
df$crop_factor     <- df_key_value$key_value[match(df$CROPDMGEXP, df_key_value$key
)]
```

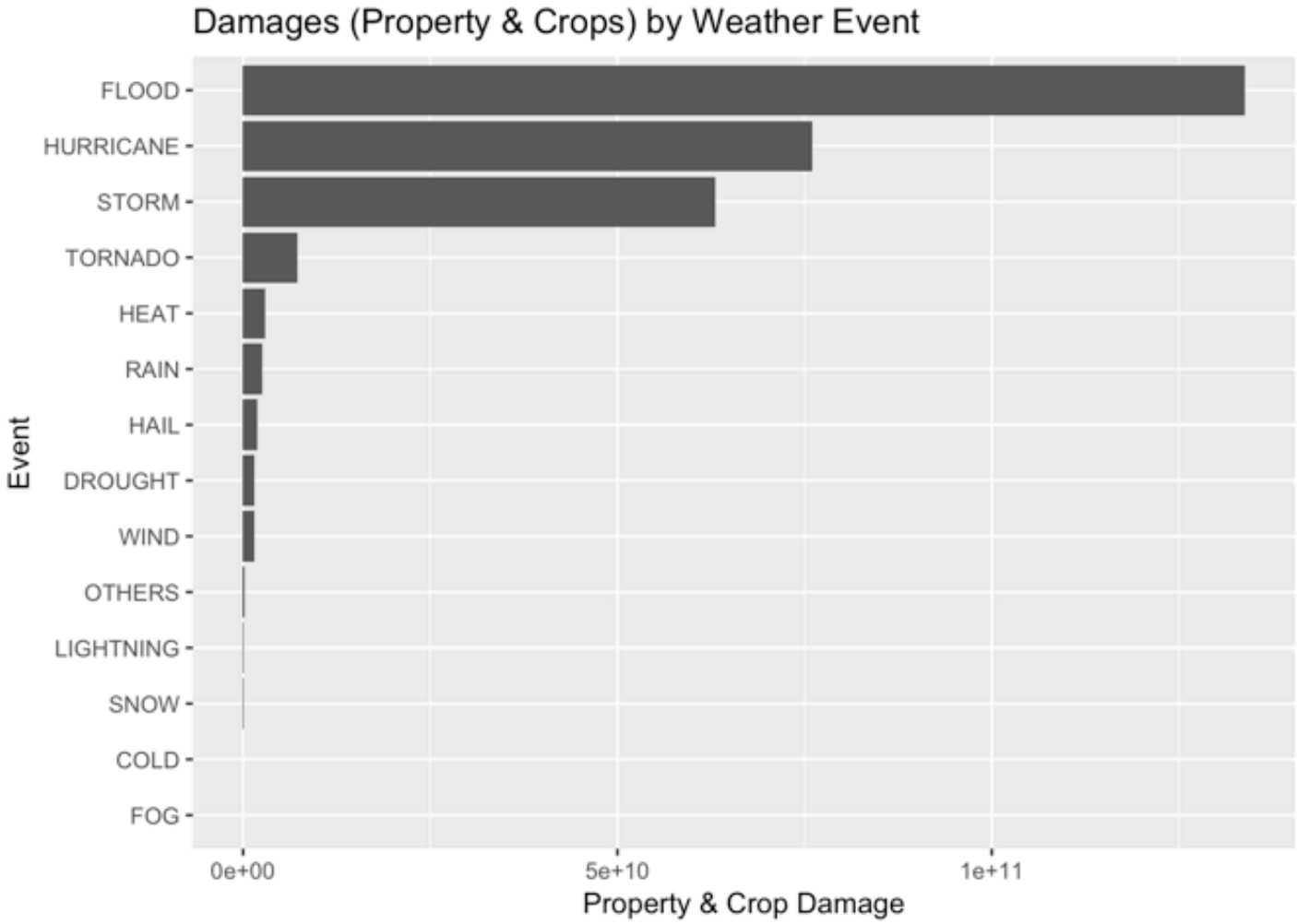
The health effects :

```
df %>%
      select(EVENT, FATALITIES, INJURIES) %>%
      group_by(EVENT) %>%
      summarise(total_fatalities = sum(FATALITIES),
                total_injuries   = sum(INJURIES),
                total_casualties = total_fatalities + total_injuries) %>%
      select(EVENT, total_casualties) %>%
      arrange(desc(total_casualties)) %>%
          ggplot(aes(x=reorder(EVENT, total_casualties), y= total_casualties)) +
              geom_bar(stat = "identity") +
              labs(x = "Event", y = "Total Casualties", title = "Casualties & Inju
ries by Weather Event") +
              coord_flip()
```



The economic consequences :

```
df %>%
      mutate(property_damage = PROPDMG * property_factor,
             crop_damage = CROPDMG * crop_factor,
             total_damage = property_damage + crop_damage) %>%
      group_by(EVENT) %>%
      summarise(damage_by_event = sum(total_damage, na.rm = T)) %>%
      arrange(desc(damage_by_event)) %>%
          ggplot(aes(x=reorder(EVENT, damage_by_event), y= damage_by_event)) +
             geom_bar(stat = "identity") +
             labs(x = "Event", y = "Property & Crop Damage", title = "Damages (Pro
perty & Crops) by Weather Event") +
             coord_flip()
```


Damages (Property & Crops) by Weather Event
FLOOD
HURRICANE
STORM
TORNADO
HEAT
RAIN
HAIL
DROUGHT
WIND
OTHERS
LIGHTNING
SNOW
COLD
FOG
Event
0e+00
5e+10
1e+11
Property & Crop Damage