Severe Weather
Casualities and Damage in the United States
1950-2011

Synopsis

Storms and severe weather often cause fatalities and injuries to people and damage to property and crops. The analysis below examines the health and economic consequences of weather events in the United States for the period 1950 to 2011 using the US National Oceanic and Atmospheric Administration's (NOAA) storm database.

Data Processing

According to the NOAA website[1],

1. Tornado: From 1950 through 1954, only tornado events were recorded.
2. Tornado, Thunderstorm Wind and Hail: From 1955 through 1992, only tornado, thunderstorm wind and hail events were keyed from the paper publications into digital data. From 1993 to 1995, only tornado, thunderstorm wind and hail events have been extracted from the Unformatted Text Files.
3. All Event Types (48 from Directive 10-1605): From 1996 to present, 48 event types are recorded as defined in NWS Directive 10-1605.

Therefore the health and economic consequences will be grouped for the following periods: 1950-1954; 1955-1992; 1993-1995; and 1996-2011.

The following libraries will be used to analyse and present the data.

```
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(ggplot2)
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

Load the data into a data frame

```
conn <- bzfile("repdata_data_StormData.csv.bz2")
df <- read.csv(conn)
```

Create a helper function to "bin" the event based on its year of occurence according to the predefined ranges discussed above.

```
year_bin <- function(year) {
  ifelse (year %in% 1950:1954, "1950-1954",
      ifelse(year %in% 1955:1992, "1955-1992",
          ifelse(year %in% 1993:1995, "1993-1995",
              ifelse(year %in% 1996:2011, "1996-2011", ""))))
}
```

Since the analysis is only interested in health and economic consequences, keep only those fields that relate to fatalities, injuries, property damage, and crop damage. The piping feature of dyplr will be used so that the event year can be binned into one of the four year ranges as well.

```
events <- df %>%
  select(BGN_DATE, EVTYPE, FATALITIES, INJURIES,PROPDMG, PROPDMGEXP,
         CROPDMG, CROPDMGEXP) %>%
  mutate(BGN_DATE = year_bin(as.numeric(strftime(as.POSIXct(BGN_DATE,
         format="%m/%d/%Y %H:%M:%S"), '%Y'))))
```

At this point, the analysis will split into health events (fatalities and injuries) and economic events (property damage and crop damage).

For health events, a helper function was created to select health event related fields, filter by year range, and total the fatalities and injuries for that year range for each event.

```
health_events <- function(df, begin, end) {
  df %>%
    select(BGN_DATE, EVTYPE, FATALITIES, INJURIES) %>%
    filter(BGN_DATE >= begin & BGN_DATE <= end) %>%
    mutate(Total = FATALITIES + INJURIES) %>%
    select(EVTYPE, Total) %>%
    group_by(EVTYPE) %>%
    summarize_each(funs(sum)) %>%
    arrange(desc(Total))
}
```

For economic events, a couple of helper functions were created. The original data set used H/h, K/k, M/m, or B/b to indicate hundreds of dollars, thousands of dollars, millions of dollars, or billions of dollars respectively. The first function creates a multiplier to convert this mixed alphanumeric damage in US dollars to damage in millions of US dollars. If any character other than one of the above mentioned occurs, then the multiplier is assigned the value of zero (0).

```
multiplier <- function(ch1) {
  multNum <- c(10^-6, 10^-4, 10^-3, 1, 10^3)
  multChar <- c("", "H", "K", "M", "B")
  ifelse(ch1 %in% multChar, multNum[match(ch1, multChar)], 0)
}
```

The second helper function selects economic related fields, filters by year range, and totals the property and crop damage for that year range for each event.

```
economic_events <- function(df, begin, end) {
  df %>%
    select(BGN_DATE, EVTYPE, PROPDMG, PROPDMGEXP, CROPDMG, CROPDMGEXP) %>%
    filter(BGN_DATE >= begin & BGN_DATE <= end) %>%
    mutate(PROPDMGEXP = toupper(PROPDMGEXP)) %>%
    mutate(PROPDMGEXP = toupper(PROPDMGEXP)) %>%
    mutate(PROPDMG = PROPDMG * multiplier(PROPDMGEXP)) %>%
    mutate(CROPDMG = CROPDMG * multiplier(CROPDMGEXP)) %>%
    mutate(Total = PROPDMG + CROPDMG) %>%
    select(EVTYPE, Total) %>%
    group_by(EVTYPE) %>%
    summarize_each(funs(sum)) %>%
    arrange(desc(Total))
}
```

Using the helper functions above, summarise the health events for each year range.

```
health_1950_1954 <-health_events(events, 1950, 1954)
health_1955_1992 <-health_events(events, 1955, 1992)
health_1993_1995 <-health_events(events, 1993, 1995)
health_1996_2011 <-health_events(events, 1996, 2011)
```

Likewise summarise the economic events for each year range.

```
economic_1950_1954 <-economic_events(events, 1950, 1954)
economic_1955_1992 <-economic_events(events, 1955, 1992)
economic_1993_1995 <-economic_events(events, 1993, 1995)
economic_1996_2011 <-economic_events(events, 1996, 2011)
```

Results

The results will be presented in a panel plot with panel representing one of the four year ranges. Only the top 5 events will be shown in each subpanel.

First, define how each subpanel looks for a range year.

```
plot_top5 <- function(event, title, ymax) {
  ggplot(data=head(event, 5),
         aes(x=reorder(EVTYPE, -Total), y=Total)) +
    geom_bar(stat = "identity") +
    theme(axis.text.x=element_text(angle=45, hjust=1, vjust=1)) +
    xlab("") +
```

```
    ylab("") +
    ylim(0, ymax) +
    ggtitle(title) +
    theme(plot.title = element_text(hjust=0.5))
}
```

So that the y-axis of each subpanel uses the same scale, determine the maximum total for health events and economic events.

```
yHealthMax <- max(health_1950_1954$Total, health_1955_1992$Total,
                  health_1993_1995$Total, health_1996_2011$Total)

yEconomicMax <- max(economic_1950_1954$Total, economic_1955_1992$Total,
                    economic_1993_1995$Total, economic_1996_2011$Total)
```

Create the four health subpanels

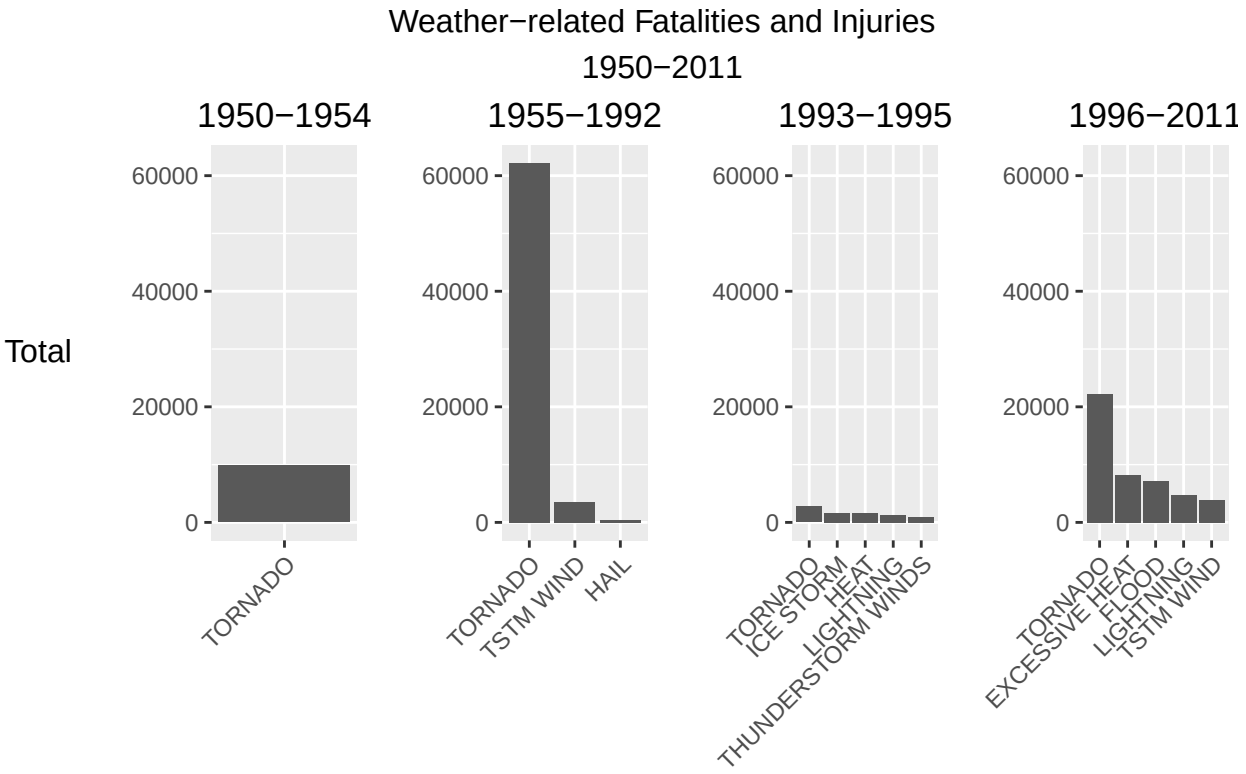
```
p1 <- plot_top5(health_1950_1954, "1950-1954", yHealthMax)
p2 <- plot_top5(health_1955_1992, "1955-1992", yHealthMax)
p3 <- plot_top5(health_1993_1995, "1993-1995", yHealthMax)
p4 <- plot_top5(health_1996_2011, "1996-2011", yHealthMax)
```

Create the four economic related subpanels

```
p5 <- plot_top5(economic_1950_1954, "1950-1954", yEconomicMax)
p6 <- plot_top5(economic_1955_1992, "1955-1992", yEconomicMax)
p7 <- plot_top5(economic_1993_1995, "1993-1995", yEconomicMax)
p8 <- plot_top5(economic_1996_2011, "1996-2011", yEconomicMax)
```

Put the health subpanels into the health panel plot

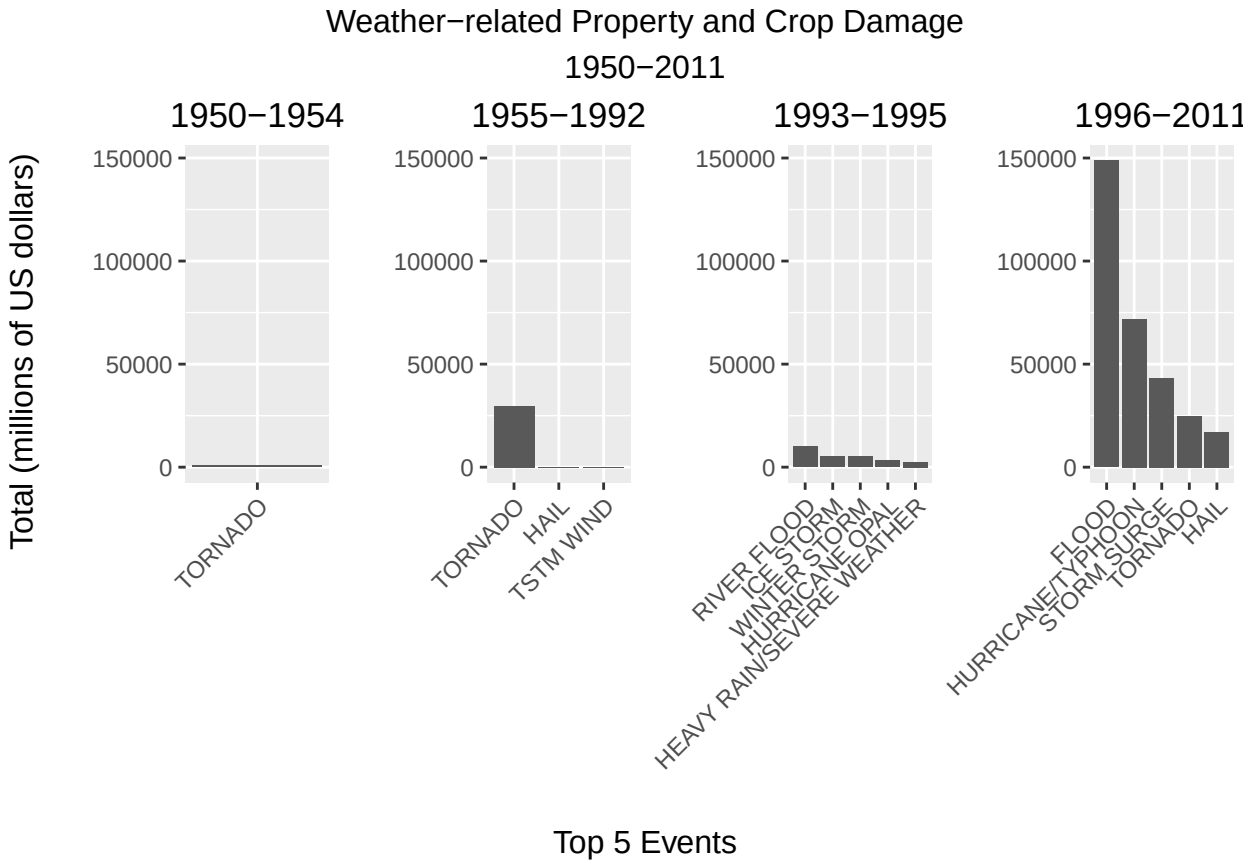
```
grid.arrange(cbind(ggplotGrob(p1),
                   ggplotGrob(p2),
                   ggplotGrob(p3),
                   ggplotGrob(p4)),
    top = textGrob("Weather-related Fatalities and Injuries\n1950-2011"),
    left = textGrob("Total", vjust=-3),
    bottom= textGrob("Top 5 Events"))
```



The health panel plot shows that tornados overall have had the greatest impact on human health.

Now put the economic subpanels into the economic panel plot.

```
grid.arrange(cbind(ggplotGrob(p5),
                   ggplotGrob(p6),
                   ggplotGrob(p7),
                   ggplotGrob(p8)),
             top = textGrob("Weather-related Property and Crop Damage\n1950-2011"),
             left = textGrob("Total (millions of US dollars)", rot=90, hjust=0.3),
             bottom= textGrob("Top 5 Events"))
```



In terms of economic costs, flooding has had the biggest dollar impact overall.

Future Work

Two areas for improving upon this analyis come to mind.

1. The analysis does NOT reconcile events that have similar names. For example, THUDERSTORM WINDS [sic], THUNDEERSTORM WINDS [sic], THUNDERESTORM WINDS [sic], THUNDERSTORM WINDS, etc are treated as separate event types.The weather events could be "binned" into the 48 event types specified in NWS Directive 10-1605.
2. Property and crop damage figures are in absolute dollars. The consumer price index could be used to adjust for inflation.

---

[1]As at 8 February 2017