Reproducible research week4 assignment

Synopsis

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.

This data analysis address the following questions:

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

Data Processing

1. load the data

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2.1
```

```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

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

```
library(ggrepel)

# load csv file
df <- read.csv(file = "repdata_data_StormData.csv")
# revise the time format
df$BGN_DATE <- as.Date(df$BGN_DATE, format="%m/%d/%Y 0:00:00")
```

2. manipulate the data

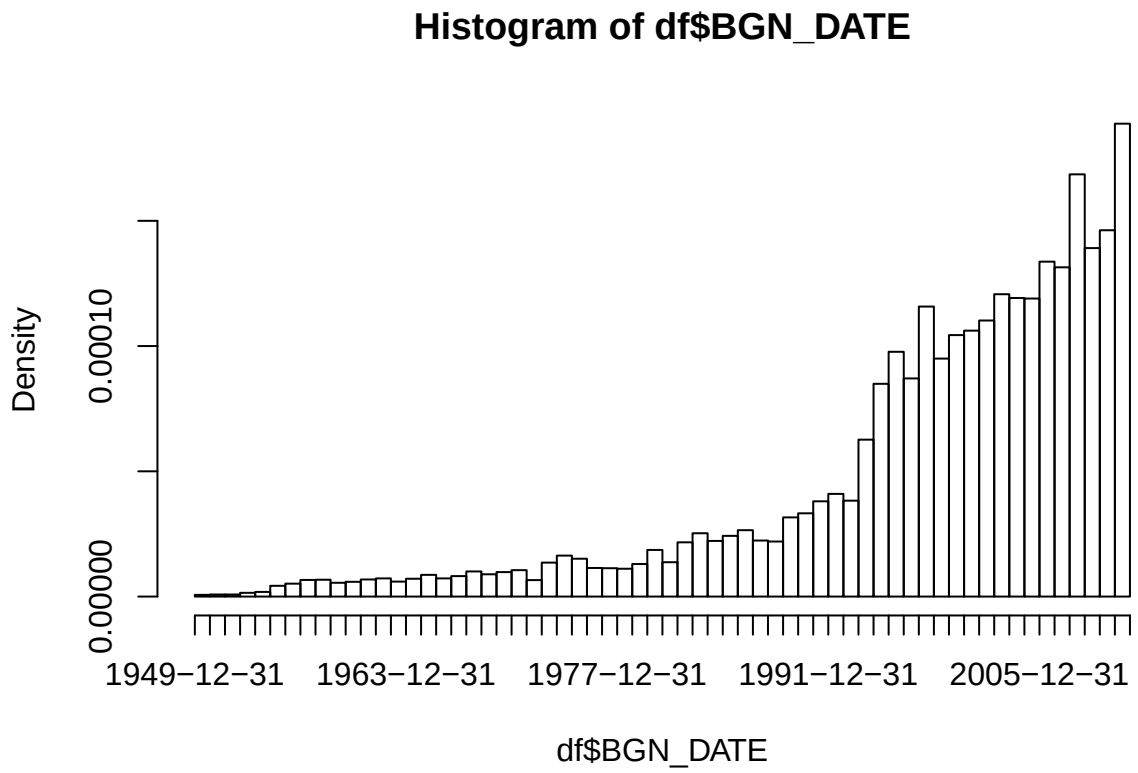
2.1 check the data

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.

```
# median of the BGN_date
median(df$BGN_DATE)
```

```
## [1] "2002-03-18"
```

```
# histgram by years
hist(df$BGN_DATE, breaks = "years")
```



2.2 selection of data

```
# pickup after 2002-01-01
df1 <- subset(df, BGN_DATE >= "2002-01-01")
head(df1)
```

```
##        STATE__   BGN_DATE    BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 447893       1 2002-06-04 01:14:00 PM       CST    123 TALLAPOOSA    AL
## 447894       1 2002-06-04 01:14:00 PM       CST    123 TALLAPOOSA    AL
## 447895       1 2002-06-04 01:20:00 PM       CST    123 TALLAPOOSA    AL
## 447896       1 2002-06-04 01:35:00 PM       CST     19   CHEROKEE    AL
## 447897       1 2002-06-04 02:00:00 PM       CST     27       CLAY    AL
## 447898       1 2002-06-04 02:10:00 PM       CST     15    CALHOUN    AL
##           EVTYPE BGN_RANGE BGN_AZI     BGN_LOCATI         END_DATE
## 447893      HAIL         0               NEWSITE 6/4/2002 0:00:00
## 447894 TSTM WIND         0               NEWSITE 6/4/2002 0:00:00
## 447895      HAIL         0         ALEXANDER CITY 6/4/2002 0:00:00
## 447896      HAIL         4       S         CENTRE 6/4/2002 0:00:00
## 447897 LIGHTNING        13       N        ASHLAND 6/4/2002 0:00:00
## 447898      HAIL         0              ANNISTON 6/4/2002 0:00:00
##           END_TIME COUNTY_END COUNTYENDN END_RANGE END_AZI     END_LOCATI
## 447893 01:14:00 PM          0         NA         0               NEWSITE
## 447894 01:14:00 PM          0         NA         0               NEWSITE
## 447895 01:20:00 PM          0         NA         0         ALEXANDER CITY
## 447896 01:35:00 PM          0         NA         4       S         CENTRE
## 447897 02:00:00 PM          0         NA        13       N        ASHLAND
## 447898 02:20:00 PM          0         NA         0              ANNISTON
##        LENGTH WIDTH  F MAG FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG
## 447893      0     0 NA 175          0        0       2          K       0
## 447894      0     0 NA  50          0        0       2          K       0
## 447895      0     0 NA 100          0        0       0          K       0
## 447896      0     0 NA  75          0        0       0          K       0
## 447897      0     0 NA   0          0        1       3          K       0
## 447898      0     0 NA 100          0        0       0          K       0
##        CROPDMGEXP WFO       STATEOFFIC ZONENAMES LATITUDE LONGITUDE
## 447893          K BMX ALABAMA, Central               3302      8546
## 447894          K BMX ALABAMA, Central               3302      8546
## 447895          K BMX ALABAMA, Central               3257      8558
## 447896          K BMX ALABAMA, Central               3406      8541
## 447897          K BMX ALABAMA, Central                  0         0
## 447898          K BMX ALABAMA, Central               3339      8550
##        LATITUDE_E LONGITUDE_
## 447893       3302       8546
## 447894       3302       8546
## 447895       3257       8558
## 447896       3406       8541
## 447897          0          0
## 447898       3339       8550
##
## 447893                                                    Golf ball size hail was reported
## 447894                                                    Golf ball size hail was reported
## 447895                                                    Golf ball size hail was reported
## 447896
## 447897 A seven year old girl was struck by lightning inside the Cheaha State Park grounds.  She suff
## 447898                                                                                        Nic
##        REFNUM
## 447893 448811
## 447894 448812
## 447895 448813
## 447896 448814
```

```
## 447897 448815
## 447898 448816
```

2.3 revise the data; from data documentation

Alphabetical characters used to signify magnitude include "K" for thousands, "M" for millions, and "B" for billions.

```
df1$PROPDMGEXP_n <- df1$PROPDMGEXP %>%
        str_replace_all(c("K" = "1000", "M" = "1000000", "B" = "1000000000"))
df1$CROPDMGEXP_n <- df1$CROPDMGEXP %>%
        str_replace_all(c("K" = "1000", "M" = "1000000", "B" = "1000000000"))
df1$PROPDMGEXP_n <- as.numeric(df1$PROPDMGEXP_n)
df1$CROPDMGEXP_n <- as.numeric(df1$CROPDMGEXP_n)
df1$PROPDMG_n <- df1$PROPDMG * df1$PROPDMGEXP_n
df1$CROPDMG_n <- df1$CROPDMG * df1$CROPDMGEXP_n
df2 <- df1[,c(8, 23, 24, 40, 41)]
head(df2)
```

```
##           EVTYPE FATALITIES INJURIES PROPDMG_n CROPDMG_n
## 447893      HAIL          0        0      2000         0
## 447894 TSTM WIND          0        0      2000         0
## 447895      HAIL          0        0         0         0
## 447896      HAIL          0        0         0         0
## 447897 LIGHTNING          0        1      3000         0
## 447898      HAIL          0        0         0         0
```

2.4 sum by event type

```
df3 <- df2 %>%
        group_by(EVTYPE) %>%
        summarise(FATALITIES_total = sum(FATALITIES, na.rm = TRUE),
                  INJURIES_total = sum(INJURIES, na.rm =  TRUE),
                  PROPDMG_total = sum(PROPDMG_n, na.rm = TRUE),
                  CROPDMG_total = sum(CROPDMG_n, na.rm =  TRUE)
                  )
head(df3)
```

```
## # A tibble: 6 x 5
##   EVTYPE        FATALITIES_total INJURIES_total PROPDMG_total CROPDMG_total
##   <fct>                    <dbl>          <dbl>         <dbl>         <dbl>
## 1 ABNORMALLY D~                0              0             0             0
## 2 ABNORMALLY W~                0              0             0             0
## 3 ASTRONOMICAL~                0              0       9425000             0
## 4 ASTRONOMICAL~                0              0        320000             0
## 5 AVALANCHE                  145            103       2722300             0
## 6 BLACK ICE                    0              0             0             0
```

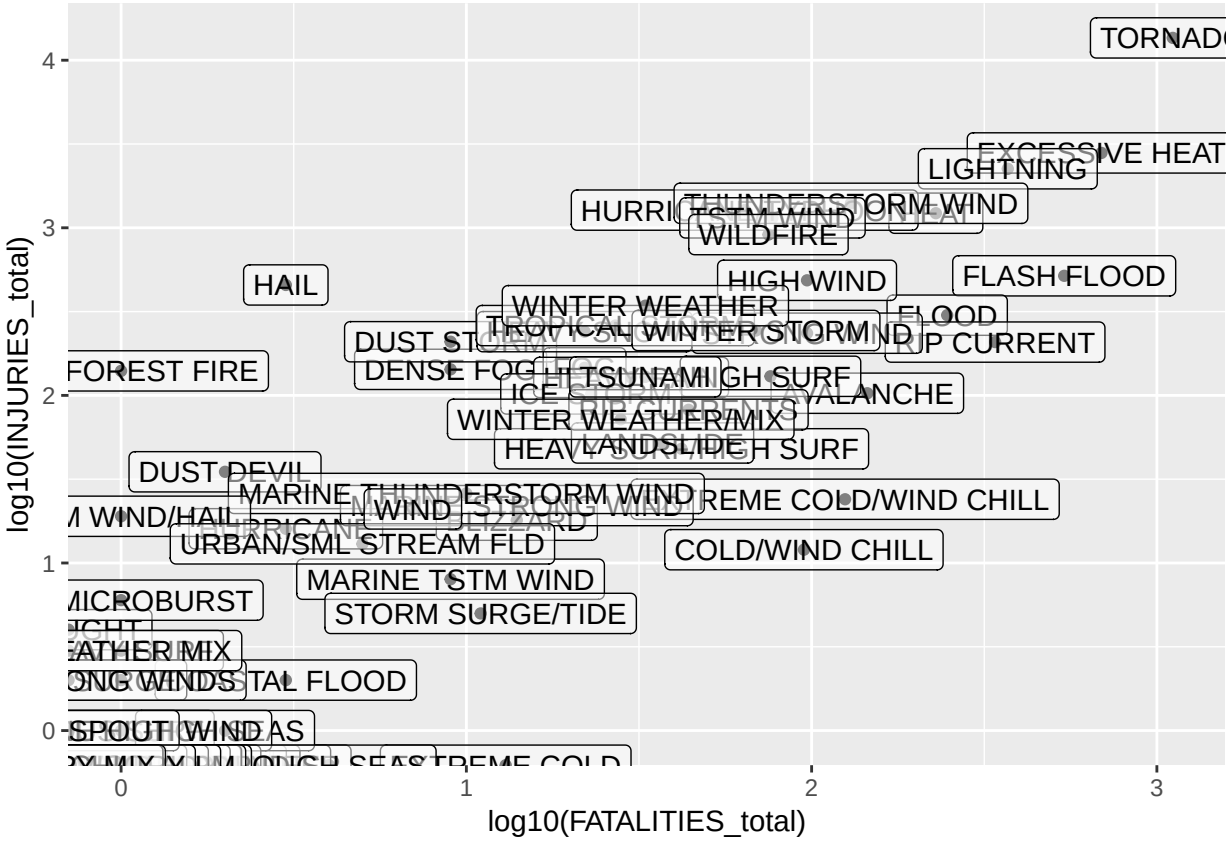
3 Results

3.1 analyze the data; damage on poplulation health

Q. Across the United States, which types of events (as indicated in the EVTYPE variable) are most harmful with respect to population health?

A. "TORNADE" is most harmful on people health.

```
g <- ggplot(df3, aes(x = log10(FATALITIES_total), y = log10(INJURIES_total), label = EVTYPE))
g + geom_point() + geom_label(alpha = 0.5)
```




```
#
df3 %>%
        arrange(desc(df3$FATALITIES_total+df3$INJURIES_total)) %>%
        head()
```

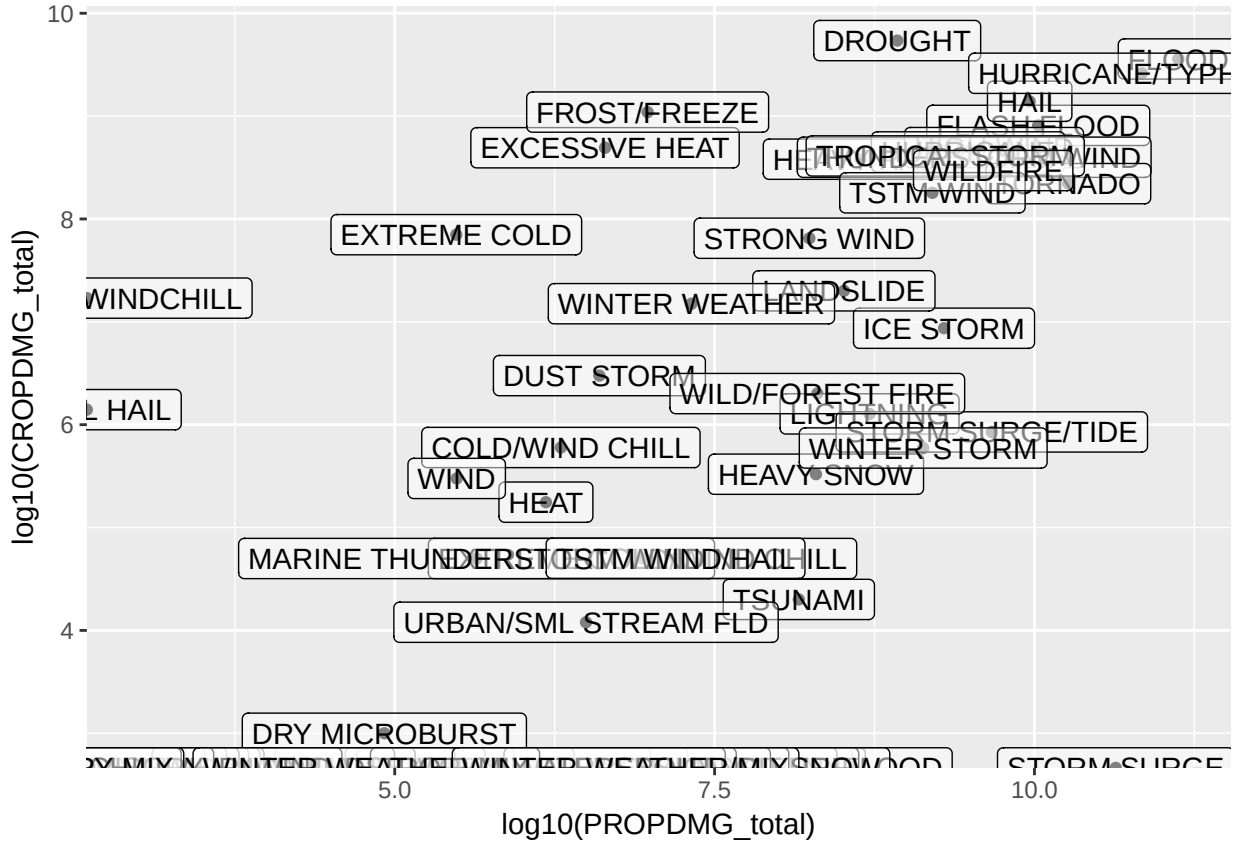
```
## # A tibble: 6 x 5
##   EVTYPE        FATALITIES_total INJURIES_total PROPDMG_total CROPDMG_total
##   <fct>                    <dbl>          <dbl>         <dbl>         <dbl>
## 1 TORNADO                   1112          13588   18406922660     220589910
## 2 EXCESSIVE HE~              691           2797       4403200     492402000
## 3 LIGHTNING                  370           2250     514320160       1272800
## 4 THUNDERSTORM~              130           1400    3382654440     398331000
## 5 HEAT                       229           1222       1520000        176500
## 6 HURRICANE/TY~               64           1275   69305840000    2607872800
```

3.2 analyze the data; ploting; damage on economics

Q. 2. Across the United States, which types of events have the greatest economic consequences?

A. "FLOOD" has the greatest impact on economics.

```
g <- ggplot(df3, aes(x = log10(PROPDMG_total), y = log10(CROPDMG_total), label = EVTYPE))
g + geom_point() + geom_label(alpha = 0.5)
```




```
#
df3 %>%
        arrange(desc(df3$PROPDMG_total+df3$CROPDMG_total)) %>%
        head()
```

```
## # A tibble: 6 x 5
##   EVTYPE        FATALITIES_total INJURIES_total PROPDMG_total CROPDMG_total
##   <fct>                    <dbl>          <dbl>         <dbl>         <dbl>
## 1 FLOOD                      247            301  133387648530    3591907400
## 2 HURRICANE/TY~               64           1275   69305840000    2607872800
## 3 STORM SURGE                  0              2   43168315000             0
## 4 TORNADO                   1112          13588   18406922660     220589910
## 5 FLASH FLOOD                539            517   10709402710     812514000
## 6 HAIL                         3            456    9174277520    1393287900
```