Reproducible Research Assignment 2

Analysis of the U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database-

In this analysis I am using the U.S. National Oceanic and Atmospheric Administration's (NOAA) data gathered between 1950 and 2011. The dataset contains information about major storms and weather events and their impact on human health and economy. The data will be used to determine which types of events coused the most catastrophic losses in human health and in the economy across the 62 years covered.

1. Data Processing

1.1 Load libraries

```
library(tidyverse)
library(lubridate)
```

1.2 Loading data into R

The following code will check if the file with the data already exists in your working directory. If it does not the file will be downloaded and unzipped.

```
if (!file.exists("storm.csv.bz2")) {

download.file("https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2","storm.csv.bz2

}
```

Load the data into R and look at the data structure

```
storm <- read.csv("storm.csv.bz2", header = TRUE)
str(storm)
```

```
## 'data.frame':    902297 obs. of  37 variables:
##  $ STATE__   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_DATE  : Factor w/ 16335 levels "1/1/1966 0:00:00",..: 6523 6523 4242 11116 2224 2224 2260 383
##  $ BGN_TIME  : Factor w/ 3608 levels "00:00:00 AM",..: 272 287 2705 1683 2584 3186 242 1683 3186 318
##  $ TIME_ZONE : Factor w/ 22 levels "ADT","AKS","AST",..: 7 7 7 7 7 7 7 7 7 7 ...
##  $ COUNTY    : num  97 3 57 89 43 77 9 123 125 57 ...
##  $ COUNTYNAME: Factor w/ 29601 levels "","5NM E OF MACKINAC BRIDGE TO PRESQUE ISLE LT MI",..: 13513 
##  $ STATE     : Factor w/ 72 levels "AK","AL","AM",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ EVTYPE    : Factor w/ 985 levels "   HIGH SURF ADVISORY",..: 834 834 834 834 834 834 834 834 834 8
##  $ BGN_RANGE : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ BGN_AZI   : Factor w/ 35 levels "","  N"," NW",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ BGN_LOCATI: Factor w/ 54429 levels "","  Christiansburg",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_DATE  : Factor w/ 6663 levels "","1/1/1993 0:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_TIME  : Factor w/ 3647 levels "","  0900CST",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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   : Factor w/ 24 levels "","E","ENE","ESE",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ END_LOCATI: Factor w/ 34506 levels "","  CANTON","  TULIA",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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: Factor w/ 19 levels "","-","?","+",..: 17 17 17 17 17 17 17 17 17 17 ...
##  $ CROPDMG   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ CROPDMGEXP: Factor w/ 9 levels "","?","0","2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ WFO       : Factor w/ 542 levels "","  CI","%SD",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ STATEOFFIC: Factor w/ 250 levels "","ALABAMA, Central",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ ZONENAMES : Factor w/ 25112 levels "","
##  $ 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   : Factor w/ 436781 levels "","\t","\t\t",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ REFNUM    : num  1 2 3 4 5 6 7 8 9 10 ...
```

1.3 Cleaning and transforming the data to facilitate the analysis

Select only variables that are neccessary for the further analysis

```
storm %>% select(REFNUM, EVTYPE, FATALITIES:CROPDMGEXP) -> storm1
```

The naming of the events and exponents contain mistakes in spelling and have variability in upper and lower case way of writing. To minimize the variablity I transform all the variables to uppercase. For the multiplier of the cost of the damage only records with valid multipliers will be kept (no multiplier, K, M, B)

```
storm1$EVTYPE <- toupper(storm1$EVTYPE)
storm1$PROPDMGEXP <- toupper(storm1$PROPDMGEXP)
storm1$CROPDMGEXP <- toupper(storm1$CROPDMGEXP)

multiplier <- c("", "K", "M", "B")
storm1 %>% filter(PROPDMGEXP %in% multiplier,
                  CROPDMGEXP %in% multiplier) -> storm2
```

The damage costs need to be unified to the same unit using the provided multiplie. K = 1000\$, M = 10^6\$, B = 10^9\$

```
storm.spl <- split(storm2, storm2$PROPDMGEXP)
```

```
as.data.frame(storm.spl[1]) %>% mutate(PROPCOST = PROPDMG) -> storm.spl0
storm.spl$K %>% mutate(PROPCOST = PROPDMG*1000) -> storm.spl1
storm.spl$M %>% mutate(PROPCOST = PROPDMG*10^6) -> storm.spl2
storm.spl$B %>% mutate(PROPCOST = PROPDMG*10^9) -> storm.spl3

storm3 <- bind_rows(storm.spl0, storm.spl1, storm.spl2, storm.spl3)

storm.spl <- split(storm3, storm3$CROPDMGEXP)

as.data.frame(storm.spl[1]) %>% mutate(CROPCOST = CROPDMG) -> storm.spl0
storm.spl$K %>% mutate(CROPCOST = CROPDMG*1000) -> storm.spl1
storm.spl$M %>% mutate(CROPCOST = CROPDMG*10^6) -> storm.spl2
storm.spl$B %>% mutate(CROPCOST = CROPDMG*10^9) -> storm.spl3

storm4 <- bind_rows(storm.spl0, storm.spl1, storm.spl2, storm.spl3)
```

The total losses to both crops and property as well as sum of injuries and fatalities need to be calculated to better understand the impact of the disasters on human life

```
storm4 %>% mutate(PROP.CROP = (PROPCOST + CROPCOST),
                  FAT.INJ = (FATALITIES + INJURIES)) -> storm4
```

2. Results

2.1 What is the type of event with the biggest impact on human health?

Calculate the time span that the data covers

```
timespan <- mdy_hms(storm$BGN_DATE)
range(timespan)
```

```
## [1] "1950-01-03 UTC" "2011-11-30 UTC"
```

```
diff <- range(timespan)[2]-range(timespan)[1]
as.duration(diff) / as.duration(years(1))
```

```
## [1] 61.90554
```

Analysed data was collected during almost 62 year period from 1950 to 2011

Calculate the total number of people injured or killed by a type of the event in the span of the 62 years.

```
storm4 %>% group_by(EVTYPE) %>% summarise(sum.fatalities = sum(FATALITIES),
                                          sum.injuries = sum(INJURIES),
                                          sum.total = sum(FAT.INJ)) %>%
        arrange(desc(sum.fatalities), desc(sum.total),
                desc(sum.injuries)) -> health
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(health, 10)
```

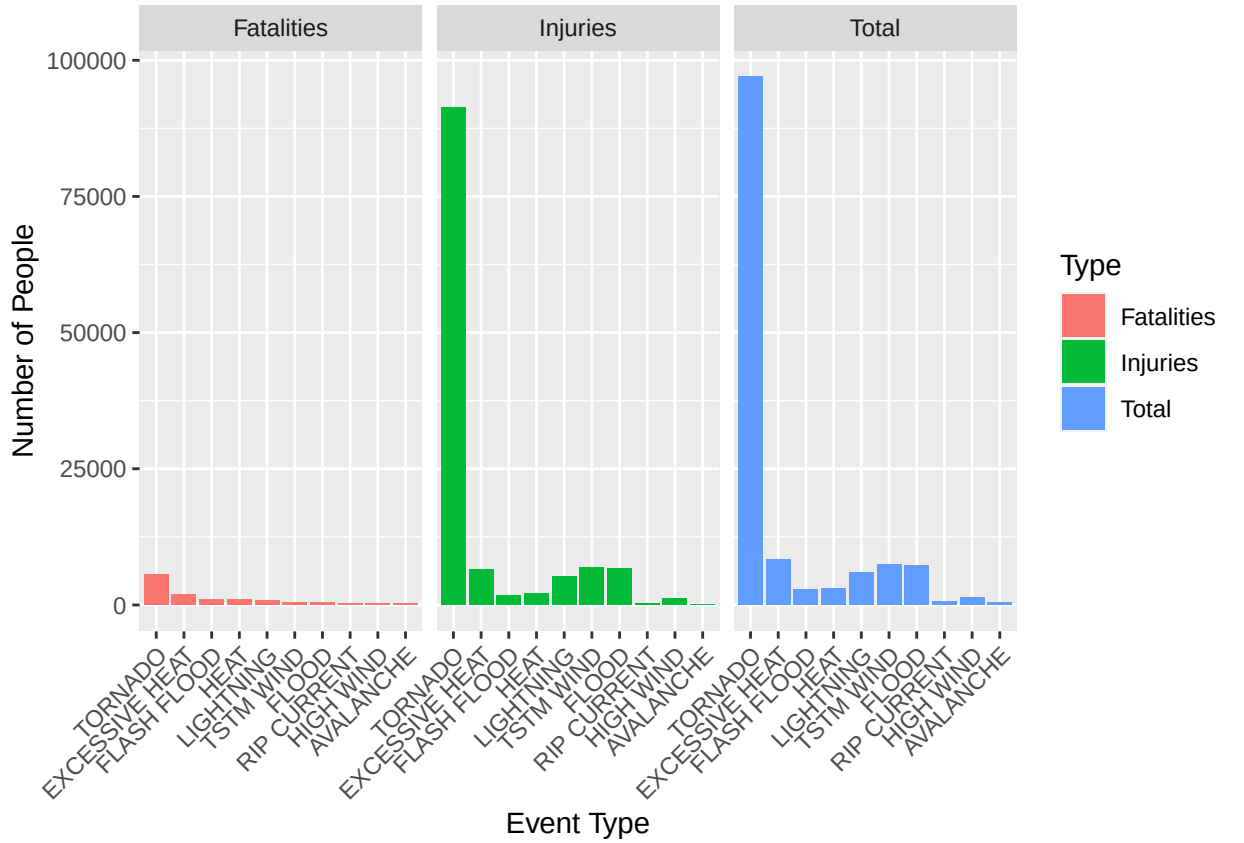
```
## # A tibble: 10 x 4
##    EVTYPE         sum.fatalities sum.injuries sum.total
##    <chr>                   <dbl>        <dbl>     <dbl>
##  1 TORNADO                  5630        91321     96951
##  2 EXCESSIVE HEAT           1903         6525      8428
##  3 FLASH FLOOD               978         1777      2755
##  4 HEAT                      937         2100      3037
##  5 LIGHTNING                 816         5230      6046
##  6 TSTM WIND                 504         6957      7461
##  7 FLOOD                     470         6789      7259
##  8 RIP CURRENT               368          232       600
##  9 HIGH WIND                 246         1137      1383
## 10 AVALANCHE                 224          170       394
```

Plot the top 10 types of events that had the highest impact of human health during the studied period.

```
names(health)[2:4] <- c("Fatalities", "Injuries", "Total")
health$EVTYPE <- factor(health$EVTYPE, levels = health$EVTYPE)

pivot_longer(head(health, 10), -EVTYPE, names_to = "Type",
             values_to = "count") -> health.l

ggplot(data = health.l, aes(EVTYPE, count)) +
        geom_col(aes(fill = Type)) +
        xlab("Event Type")+
        ylab("Number of People")+
        theme(axis.text.x = element_text(angle = 45, hjust = 1))+
        facet_wrap(.~ Type)
```



Event most harmful to the human health in terms of both fatalities and injuries in the 62 years during which the data were collected were tornados.

2.1 What is the type of event with the biggest economic consequences?

Calculate the total losses by the event type in the 62 years studdied

```
storm4 %>% group_by(EVTYPE) %>% summarise(sum.prop = sum(PROPCOST),
                                          sum.crop = sum(CROPCOST),
                                          sum.total = sum(PROP.CROP)) %>%
        arrange(desc(sum.total)) -> losses
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
head(losses, 10)
```

```
## # A tibble: 10 x 4
##    EVTYPE               sum.prop    sum.crop    sum.total
##    <chr>                   <dbl>       <dbl>        <dbl>
##  1 FLOOD             144657709807  5661968450 150319678257
##  2 HURRICANE/TYPHOON  69305840000  2607872800  71913712800
##  3 TORNADO            56936985483   364950110  57301935593
##  4 STORM SURGE        43323536000        5000  43323541000
##  5 HAIL               15732261777  3000954453  18733216230
```

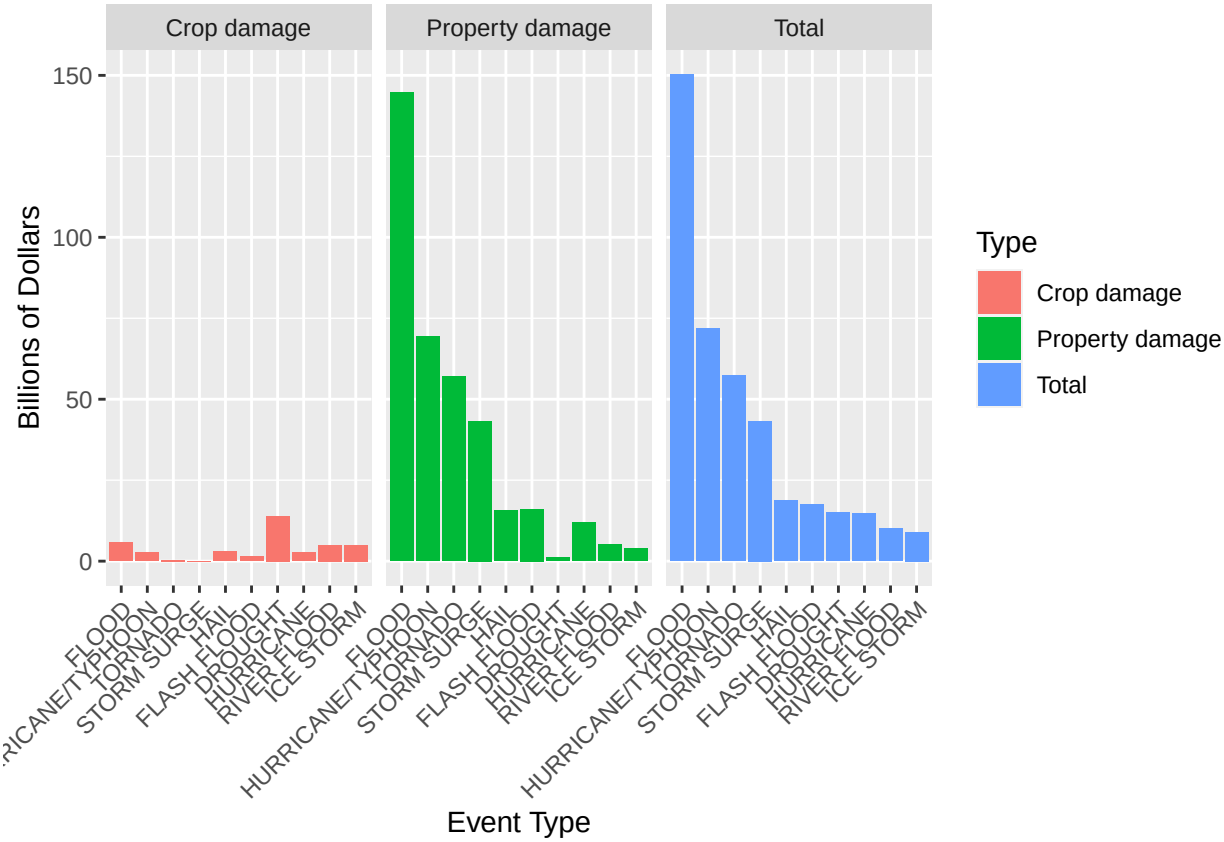
```
##  6 FLASH FLOOD         16140811717  1420727100  17561538817
##  7 DROUGHT              1046106000 13972566000  15018672000
##  8 HURRICANE           11868319010  2741910000  14610229010
##  9 RIVER FLOOD          5118945500  5029459000  10148404500
## 10 ICE STORM            3944927810  5022110000   8967037810
```

Plot the top 10 types of events that caused the highest losses during the studied period.

```
names(losses)[2:4] <- c("Property damage", "Crop damage", "Total")

losses$EVTYPE <- factor(losses$EVTYPE, levels = losses$EVTYPE)

pivot_longer(head(losses, 10), -EVTYPE, names_to = "Type",
             values_to = "count") -> losses.l
ggplot(data = losses.l, aes(EVTYPE, count/10^9)) +
        geom_col(aes(fill = Type)) +
        xlab("Event Type")+
        ylab("Billions of Dollars")+
        theme(axis.text.x = element_text(angle = 45, hjust = 1))+
        facet_wrap(.~ Type)
```



Event causing the highest financial losses in the 62 years during which the data were collected were floods. However, events with causing the highest losses in crops were droughts.

3. Take home message

- **Events most harmful to human health were tornados**
- **Events most harmful to economy were floods. However droughts caused most damage to crops**