USA Storm Data Analysis

Introduction:

Storms and other severe weather events can cause both public health and economic problems for communities and municipalities. Many severe events can result in fatalities, injuries, and property damage, and preventing such outcomes to the extent possible is a key concern.

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.

Synopsis

In this project we will address 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?

The data for this assignment come in the form of a comma-separated-value file compressed via the bzip2 algorithm to reduce its size. We can download the file from the web site: Storm 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.

Data Loading:

The dataset is stored in a comma-separated-value (CSV) file, and there are 902297 observations in this dataset. Total 37 variables were included in this dataset.

We will select only 9 relevant variables from the original dataset for our analysis. To optimize memory use, we will remove the original dataset. We will proceed with the data frame storm data for further analysis.

Data processing:

```
stormdata <- select(data, COUNTY, STATE, EVTYPE, FATALITIES, INJURIES, PROPDMG, PROPDMGEXP, CROPDMG, CR
rm(data)
glimpse(stormdata)
```

```
## Rows: 902,297
## Columns: 9
## $ COUNTY     <dbl> 97, 3, 57, 89, 43, 77, 9, 123, 125, 57, 43, 9, 73, 49, 107,~
## $ STATE      <chr> "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL", "AL",~
## $ EVTYPE     <chr> "TORNADO", "TORNADO", "TORNADO", "TORNADO", "TORNADO", "TOR~
## $ FATALITIES <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 4, 0, 0, 0, 0,~
## $ INJURIES   <dbl> 15, 0, 2, 2, 2, 6, 1, 0, 14, 0, 3, 3, 26, 12, 6, 50, 2, 0, ~
## $ PROPDMG    <dbl> 25.0, 2.5, 25.0, 2.5, 2.5, 2.5, 2.5, 2.5, 25.0, 25.0, 2.5, ~
## $ PROPDMGEXP <chr> "K", "K", "K", "K", "K", "K", "K", "K", "K", "K", "M", "M",~
## $ CROPDMG    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
```

```
## $ CROPDMGEXP <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", "",~
```

In the new datafrma named stormdata consists of 9 variable: COUNTY: harmful event county location

STATE: harmful event state location

EVETYPE : disastrous event type

FATALITIES: amount of fatalities per event

INJURIES : amount of injuries per event

PROPDMG : property damage amount

PROPDMGEXP: property damage in exponents

CROPDMG : crop damage amount

CROPDMGEXP: crop damage in exponents

Here is the summary of stormdata dataframe.

```
summary(stormdata)
```

```
##      COUNTY         STATE              EVTYPE            FATALITIES
##  Min.   :  0.0   Length:902297      Length:902297      Min.   :  0.0000
##  1st Qu.: 31.0   Class :character   Class :character   1st Qu.:  0.0000
##  Median : 75.0   Mode  :character   Mode  :character   Median :  0.0000
##  Mean   :100.6                                         Mean   :  0.0168
##  3rd Qu.:131.0                                         3rd Qu.:  0.0000
##  Max.   :873.0                                         Max.   :583.0000
##     INJURIES            PROPDMG         PROPDMGEXP           CROPDMG
##  Min.   :   0.0000   Min.   :   0.00   Length:902297      Min.   :  0.000
##  1st Qu.:   0.0000   1st Qu.:   0.00   Class :character   1st Qu.:  0.000
##  Median :   0.0000   Median :   0.00   Mode  :character   Median :  0.000
##  Mean   :   0.1557   Mean   :  12.06                      Mean   :  1.527
##  3rd Qu.:   0.0000   3rd Qu.:   0.50                      3rd Qu.:  0.000
##  Max.   :1700.0000   Max.   :5000.00                      Max.   :990.000
##   CROPDMGEXP
##  Length:902297
##  Class :character
##  Mode  :character
##
##
##
```

Most harmful events with respect to population health:

Here we calculate the number of harmful events and their names. We trim additional space from the name of destructive events. Then we calculate the top 10 frequencies of harmful events.

```
length(unique(stormdata$EVTYPE))
```

```
## [1] 985
```

```
#unique(stormdata$EVTYPE)
stormdata$EVTYPE<-str_trim(stormdata$EVTYPE,"left")
stormdata$EVTYPE<-str_trim(stormdata$EVTYPE,"right")
head(sort(table(stormdata$EVTYPE),decreasing=T),10)
```

```
##
```

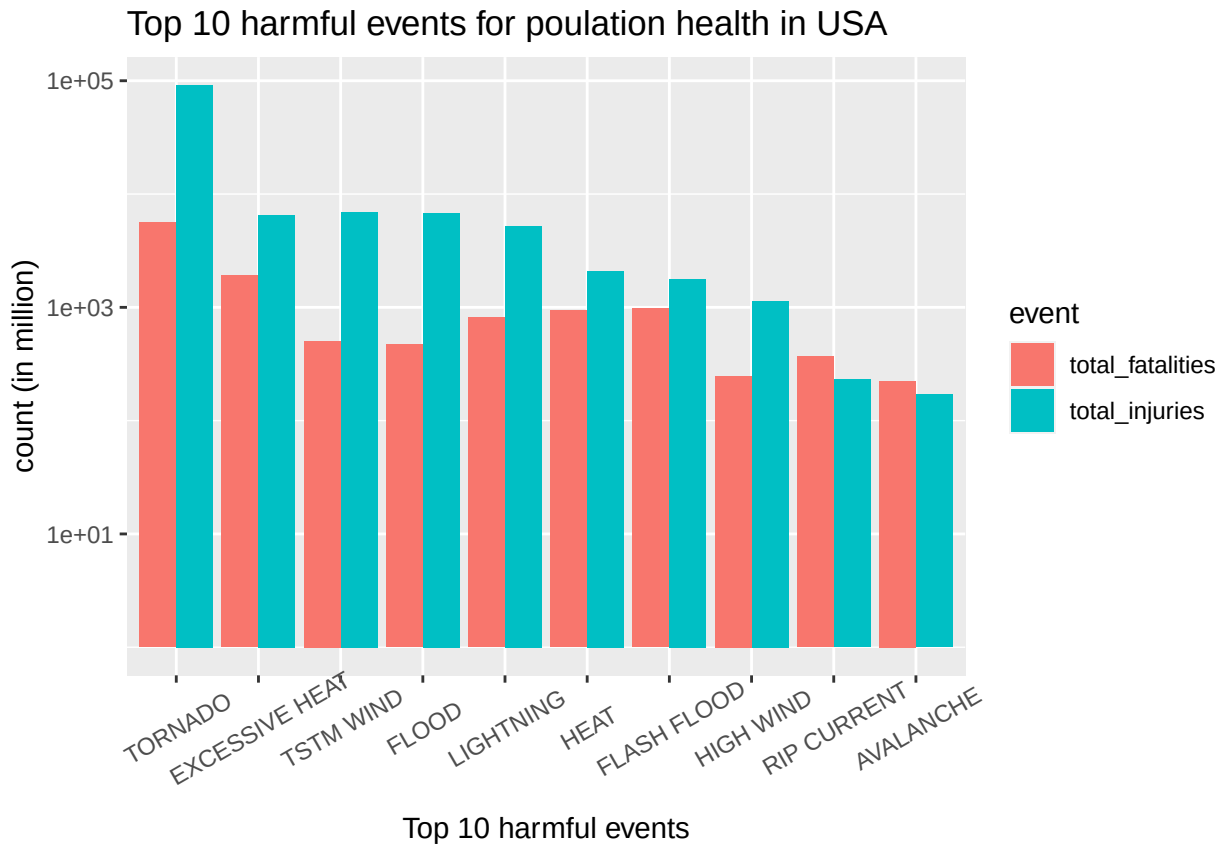
```
##              HAIL         TSTM WIND  THUNDERSTORM WIND           TORNADO
##            288661            219944              82563             60652
##       FLASH FLOOD             FLOOD THUNDERSTORM WINDS         HIGH WIND
##             54278             25326              20843             20212
##         LIGHTNING        HEAVY SNOW
##             15755             15708
```

We consider high fatality and high injury-causing events as the most harmful events concerning population health.

```
most_harmful_events<-stormdata %>%
  select(EVTYPE,FATALITIES,INJURIES)%>%
  group_by(EVTYPE) %>%
  summarize(total_fatalities = sum(FATALITIES), total_injuries = sum(INJURIES), .groups='drop')%>%
  filter(total_fatalities >0 | total_injuries >0) %>%
  arrange(desc(total_fatalities),desc(total_injuries)) %>% slice(1:10) %>%
pivot_longer(total_fatalities:total_injuries, names_to = "event")
```

The bar plot of most harmful events concerning population health in the log scale showed that tornadoes caused the highest fatalities and injuries.

```
ggplot(most_harmful_events, aes(fill=event, y=value, x=reorder(EVTYPE,-value),value)) +
geom_bar(position="dodge", stat="identity")+scale_y_log10()+
  theme(axis.text.x = element_text(angle = 30,vjust = 0.7))+xlab("Top 10 harmful events")+
  ylab("count (in million)")+ggtitle("Top 10 harmful events for poulation health in USA")
```



Across the United States, greatest economic consequence causing events:

To study the most significant economic consequence-causing events across the USA, we need to calculate the cost of property damage and crop damage caused by natural calamities.

```
unique(stormdata$PROPDMGEXP)
```

```
##  [1] "K" "M" ""  "B" "m" "+" "0" "5" "6" "?" "4" "2" "3" "h" "7" "H" "-" "1" "8"
```

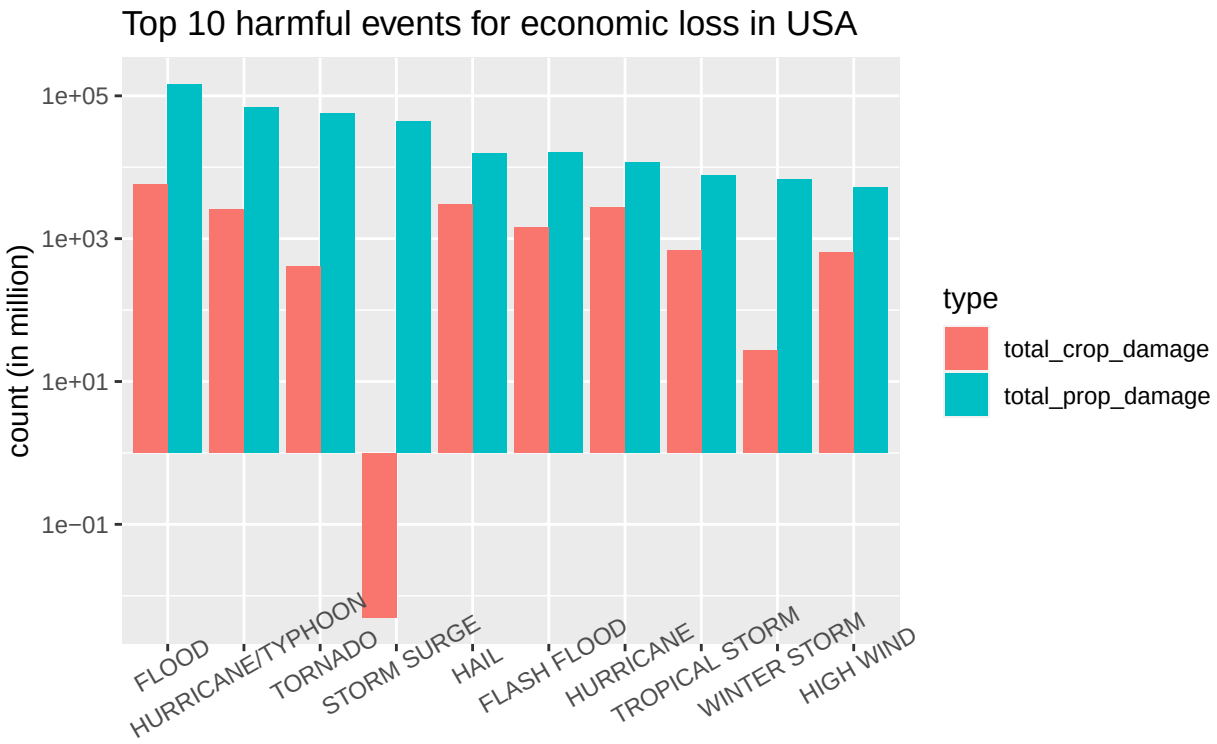
```
unique(stormdata$CROPDMGEXP)
```

```
## [1] ""  "M" "K" "m" "B" "?" "0" "k" "2"
```

```
economic_damage<-stormdata %>%
  select(EVTYPE,PROPDMG,PROPDMGEXP,CROPDMG,CROPDMGEXP) %>%
  mutate(PROPDMG_value=case_when(PROPDMGEXP=="H"~ PROPDMG*1E2
                                 ,PROPDMGEXP=="K"~PROPDMG*1E3
                                 ,PROPDMGEXP=="M"~PROPDMG*1E6
                                 ,PROPDMGEXP=="B"~PROPDMG*1E9)) %>%
  mutate(CROPDMG_value=case_when(CROPDMGEXP=="H"~ CROPDMG*1E2
                                 ,CROPDMGEXP=="K"~CROPDMG*1E3
                                 ,CROPDMGEXP=="M"~CROPDMG*1E6
                                 ,CROPDMGEXP=="B"~CROPDMG*1E9)) %>%
  group_by(EVTYPE) %>%
  summarize(total_prop_damage  = sum(PROPDMG_value,na.rm = T)/1E6,
            total_crop_damage = sum(CROPDMG_value,na.rm = T)/1E6, .groups='drop')%>%
  arrange(desc(total_prop_damage),desc(total_crop_damage)) %>%
  slice(1:10) %>%
  gather(key = type, value = value, total_prop_damage, total_crop_damage)
```

In the bar plot (y axis is in log scale), floods are causing most property damage and crop damage, thus resulting in the most significant economic consequences.

```
ggplot(economic_damage, aes(fill=type, y=value, x=reorder(EVTYPE,-value),value)) +
  geom_bar(position="dodge", stat="identity")+scale_y_log10()+
  theme(axis.text.x = element_text(angle = 30))+xlab("Top 10 harmful events")+
  ylab("count (in million)")+ggtitle("Top 10 harmful events for economic loss in USA")
```



Conclusion:

Based on the analysis, most fatalities and injuries are caused by tornadoes. To lower the population loss caused by tornadoes, we should focus on developing better warning systems. We should build better water management systems to minimize economic losses like property damage and crop damage caused by floods.