Reproducible Research Project 2

Synopsis

The basic goal of this assignment is to explore the NOAA Storm Database and answer two questions: which types of events are most harmful to population health and which types of events have the greatest economic consequences. From the data set, we found out that **TORNADO** has the largest impact on damaging both population and economy.

Data Processing

Reading the raw data

```
storm<- read.csv("/users/andrewhu/desktop/storm.csv")
```

Previewing the structure of the data

```
head(storm)
```

```
##   STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STATE
## 1       1  4/18/1950 0:00:00     0130       CST     97     MOBILE    AL
## 2       1  4/18/1950 0:00:00     0145       CST      3    BALDWIN    AL
## 3       1  2/20/1951 0:00:00     1600       CST     57    FAYETTE    AL
## 4       1   6/8/1951 0:00:00     0900       CST     89    MADISON    AL
## 5       1 11/15/1951 0:00:00     1500       CST     43    CULLMAN    AL
## 6       1 11/15/1951 0:00:00     2000       CST     77 LAUDERDALE    AL
##    EVTYPE BGN_RANGE BGN_AZI BGN_LOCATI END_DATE END_TIME COUNTY_END
## 1 TORNADO         0                                               0
## 2 TORNADO         0                                               0
## 3 TORNADO         0                                               0
## 4 TORNADO         0                                               0
## 5 TORNADO         0                                               0
## 6 TORNADO         0                                               0
##   COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALITIES
## 1         NA         0                      14.0   100 3   0          0
## 2         NA         0                       2.0   150 2   0          0
## 3         NA         0                       0.1   123 2   0          0
## 4         NA         0                       0.0   100 2   0          0
## 5         NA         0                       0.0   150 2   0          0
## 6         NA         0                       1.5   177 2   0          0
##   INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONENAMES
## 1       15    25.0          K       0
## 2        0     2.5          K       0
## 3        2    25.0          K       0
## 4        2     2.5          K       0
## 5        2     2.5          K       0
## 6        6     2.5          K       0
##   LATITUDE LONGITUDE LATITUDE_E LONGITUDE_ REMARKS REFNUM
## 1     3040      8812       3051       8806              1
## 2     3042      8755          0          0              2
## 3     3340      8742          0          0              3
## 4     3458      8626          0          0              4
## 5     3412      8642          0          0              5
## 6     3450      8748          0          0              6
```

```
dim(storm)
```

```
## [1] 902297     37
```

Finding the variables we are interested

The columns we are intersted related to the **harmfulness of Population**, are the "Fatalities" and "Injuries". Here we take a look of their summaries:

```
summary(storm$FATALITIES)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000   0.0000   0.0000   0.0168   0.0000 583.0000
```

```
summary(storm$INJURIES)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0000    0.0000    0.0000    0.1557    0.0000 1700.0000
```

Create a variable called **dmg_pop** indicating the damage of the population, combining the fatalities and injuries.

```
storm$dmg_pop = storm$FATALITIES + storm$INJURIES
```

Calculate the dmg_pop by each type of events and return a new data frame

```
popdamage <- aggregate(dmg_pop~EVTYPE, data=storm, sum)
```

Simply taking a look of the new data frame we just create, we found that there are a lot of EVTYPE, and many of the EVTYPE contain 0 dmg_pop

```
summary(popdamage)
```

```
##                     EVTYPE       dmg_pop
##     HIGH SURF ADVISORY:  1   Min.   :    0
##   COASTAL FLOOD       :  1   1st Qu.:    0
##   FLASH FLOOD         :  1   Median :    0
##   LIGHTNING           :  1   Mean   :  158
##   TSTM WIND           :  1   3rd Qu.:    0
##   TSTM WIND (G45)     :  1   Max.   :96979
##  (Other)              :979
```

```
head(popdamage)
```

```
##                   EVTYPE dmg_pop
## 1    HIGH SURF ADVISORY       0
## 2         COASTAL FLOOD       0
## 3           FLASH FLOOD       0
## 4             LIGHTNING       0
## 5             TSTM WIND       0
## 6       TSTM WIND (G45)       0
```

Hence, we need to "summary" the popdamage data frame. We can subset a data frame which contains top 5 damages for each EVTYPE.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.1
```

```
##
## 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
```

```
popdmgtop5 <- popdamage %>% arrange(desc(dmg_pop)) %>% slice(1:5)
```

Now, the variables we are interested for population damage related are processed finished. Let's take a loot at economic damage-related variables, which are "PROPDMG" and "CROPDMG".

```
summary(storm$PROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    0.00   12.06    0.50 5000.00
```

```
summary(storm$CROPDMG)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.527   0.000 990.000
```

Then, just as the steps we created for the damage for population above, we simply create a variable indicating the total impact for the economic damage, combining the two variables.

```
storm$dmg_eco <- storm$PROPDMG + storm$CROPDMG
```

Now, calculate the sum of dmg_eco for each event type and return a data frame

```
ecodamage <- aggregate(dmg_eco ~ EVTYPE, data=storm, sum)
```

Simply take a look at the new data frame for eco damage:

```
head(ecodamage)
```

```
##                  EVTYPE dmg_eco
## 1   HIGH SURF ADVISORY     200
## 2        COASTAL FLOOD       0
## 3          FLASH FLOOD      50
## 4            LIGHTNING       0
## 5            TSTM WIND     108
## 6      TSTM WIND (G45)       8
```

Filter the ecodamage for containing top 5 damges of EVTYPE only:

```
library(dplyr)
ecodmgtop5 <- ecodamage %>% arrange(desc(dmg_eco)) %>% slice(1:5)
```

Results

Now, simply printing out the popdmgtop5 and ecodmgtop5, we can have an idea of which EVTYPE has the largest impact on the population and economy:

```
popdmgtop5
```

```
##            EVTYPE dmg_pop
## 1         TORNADO   96979
## 2  EXCESSIVE HEAT    8428
## 3       TSTM WIND    7461
## 4           FLOOD    7259
## 5       LIGHTNING    6046
```

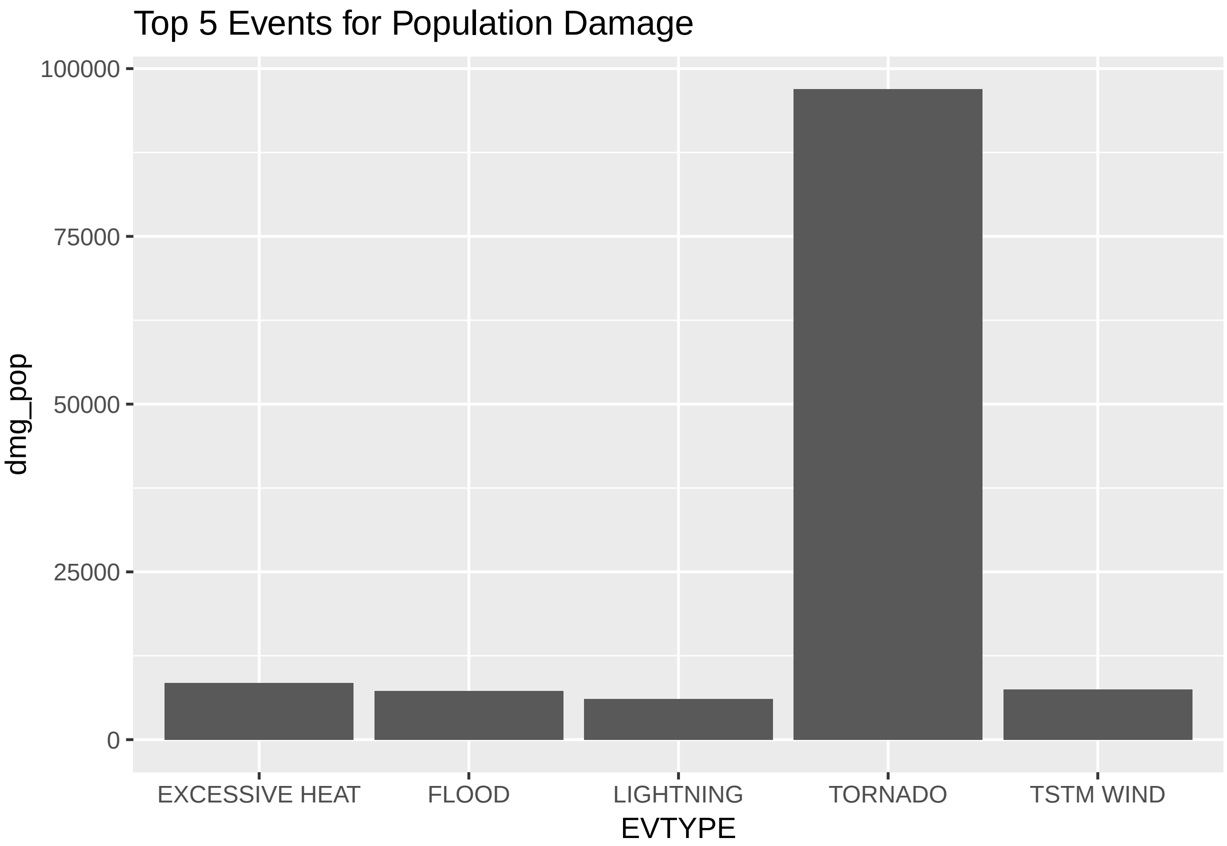
```
ecodmgtop5
```

```
##         EVTYPE dmg_eco
## 1      TORNADO 3312277
## 2  FLASH FLOOD 1599325
## 3    TSTM WIND 1445168
## 4         HAIL 1268290
## 5        FLOOD 1067976
```

In addition, let's do some plots.

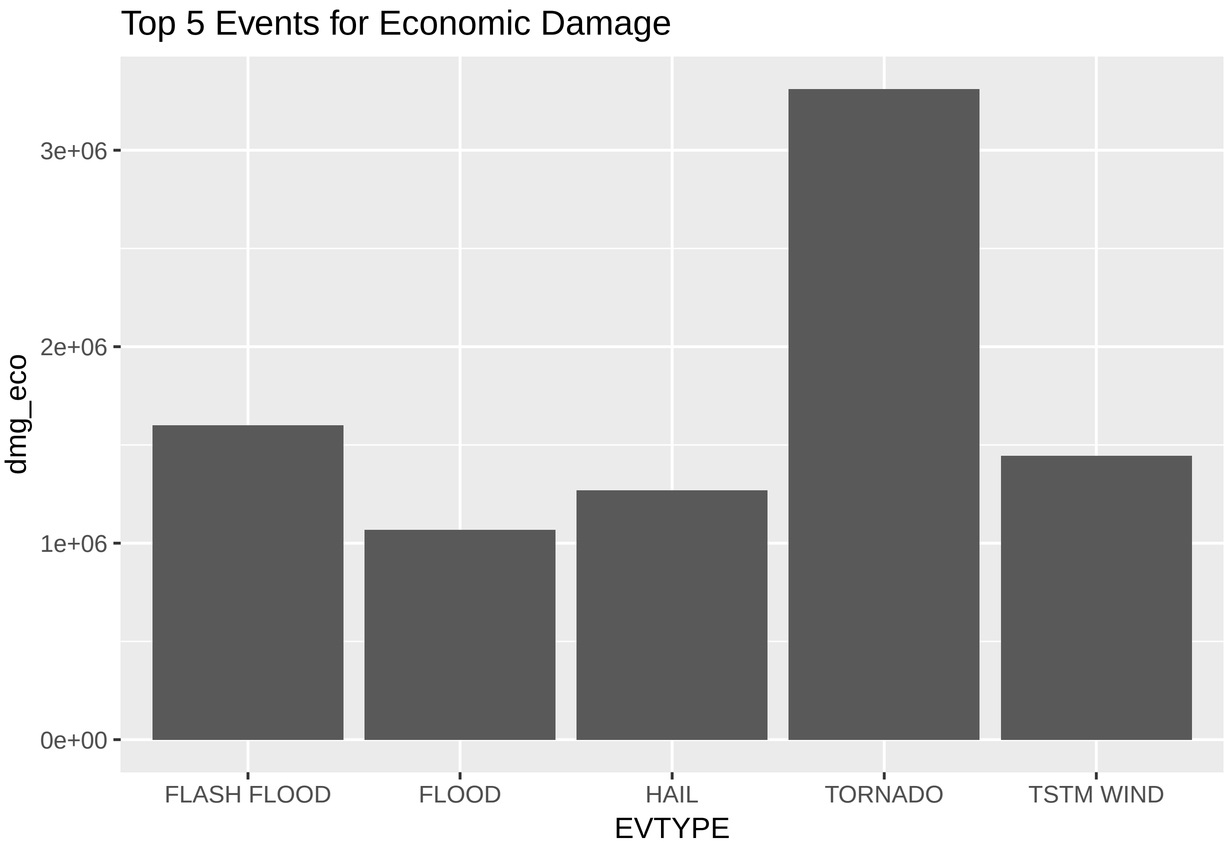
Population Damage:

```
library(ggplot2)
##ggplot
#base
g<- ggplot(popdmgtop5, aes(x=EVTYPE, y=dmg_pop))
#
g + geom_bar(stat= "identity") + labs(title= "Top 5 Events for Population Damage")
```



Economic Damage:

```
##ggplot
#base
g<- ggplot(ecodmgtop5, aes(x=EVTYPE, y=dmg_eco))
#
g + geom_bar(stat= "identity") + labs(title= "Top 5 Events for Economic Damage")
```


Top 5 Events for Economic Damage
dmg_eco
3e+06
2e+06
1e+06
0e+00
FLASH FLOOD
FLOOD
HAIL
TORNADO
TSTM WIND
EVTYPE