Storms and Severe Weather Events causes Public Health and Economic Problems

Samarjit Roy

March 6, 2016

Introducton

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.

Data

The data for this analysis comes in the form of a comma-separated-value file compressed via the bzip2 algorithm from the following link :

- Storm Data

```r
#Download Storm Data
DataURL="https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
downloadFileName <- "repdata-data-StormData.csv.bz2"
if (!file.exists(downloadFileName))
{
  download.file(DataURL, dest=downloadFileName, method="libcurl",quiet = TRUE)
}
# if zip file exists, unzip with overwrite
stormFileName <- "repdata-data-StormData.csv"
if (file.exists(downloadFileName) && !file.exists(stormFileName))
{
    bunzip2(downloadFileName, stormFileName, remove = FALSE, skip = TRUE)
}
stormDataset <- read.csv(stormFileName)
```

sample Data for Storm Data:

```r
head(stormDataset)
```

```
##   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
```

Data Transformations

Data Scope

Due to changes in the data collection and processing procedures over time, there are unique periods of records available depending on the event type. NOAA's National Weather Service (NWS) has classified data into the following Event Types:

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 we are selecting only the data that has been collected from the 1996-2011 time period, and the data set was filtered down to contain events that happened on or after Jan 1, 1996. A total of 653530 records were retained, ranging from Jan 1, 1996 to November 30, 2011.

Formatting Date Columns Let's convert the BGN_DATE column to the POSIXct format. Also add a new column, EventYear, for Year of the Event.

```
stormDataset$BGN_DATE <- as.POSIXct(stormDataset$BGN_DATE, format="%m/%d/%Y %H:%M:%S")
stormDataset$EventYear <- format(stormDataset$BGN_DATE, "%Y")
```

```
stormDataset <- filter(stormDataset, BGN_DATE >= as.POSIXct('1/1/1996', format="%m/%d/%Y"))
#head(stormDataset)
```

Restrict Data range to Jan 1,1996 - November 30, 2011

Selecting only the colums required

Storm Data has lots of information, which we do not need for this analysis. Let's select only the colums required.

```
stormDataset <- select(stormDataset, EVTYPE,FATALITIES, INJURIES, PROPDMG, PROPDMGEXP,CROPDMG, CROPDMGEX
stormDataset <- filter(stormDataset, FATALITIES>0 | INJURIES>0 | PROPDMG>0 | CROPDMG>0)
head(stormDataset)
```

```
##         EVTYPE FATALITIES INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP
## 1 WINTER STORM          0        0     380          K      38          K
## 2      TORNADO          0        0     100          K       0
## 3    TSTM WIND          0        0       3          K       0
## 4    TSTM WIND          0        0       5          K       0
## 5    TSTM WIND          0        0       2          K       0
## 6    HIGH WIND          0        0     400          K       0
##           STATEOFFIC EventYear
## 1   ALABAMA, Central      1996
## 2 ALABAMA, Southeast      1996
## 3 ALABAMA, Southeast      1996
## 4 ALABAMA, Southeast      1996
## 5 ALABAMA, Southeast      1996
## 6   ALABAMA, Central      1996
```

Removing Monthy And Yearly Data rows

```
stormDataset$EVTYPE <- str_trim(toupper(stormDataset$EVTYPE))
stormDataset <- filter(stormDataset, !grepl('SUMMARY|MONTHLY', EVTYPE))
```

Event type

Event Type is an important factor for our Storm Data Analysis. We need to verify the accuracy of the Event type.

```
EventTypes <- sort(unique(stormDataset$EVTYPE))
head(EventTypes,20)
```

```
##  [1] "AGRICULTURAL FREEZE"       "ASTRONOMICAL HIGH TIDE"
##  [3] "ASTRONOMICAL LOW TIDE"     "AVALANCHE"
##  [5] "BEACH EROSION"             "BLACK ICE"
##  [7] "BLIZZARD"                  "BLOWING DUST"
##  [9] "BLOWING SNOW"              "BRUSH FIRE"
## [11] "COASTAL  FLOODING/EROSION" "COASTAL EROSION"
## [13] "COASTAL FLOOD"             "COASTAL FLOODING"
## [15] "COASTAL FLOODING/EROSION"  "COASTAL STORM"
## [17] "COASTALSTORM"              "COLD"
## [19] "COLD AND SNOW"             "COLD TEMPERATURE"
```

```r
tail(EventTypes)
```

```
## [1] "WINDS"               "WINTER STORM"       "WINTER WEATHER"
## [4] "WINTER WEATHER MIX" "WINTER WEATHER/MIX" "WINTRY MIX"
```

NOAA's National Weather Service (NWS) has also clearly said that there are only 48 Event Types which is not the same as what we see in storm data. The 48 Event Types also defined in the Storm Data Documentation. Looks like we need to clean the Event Types. To come up with a reasonable number of Event Types, We are using the Hierarchical Cluster Analysis method.

```r
EventTypes <- sort(unique(stormDataset$EVTYPE))
distanceEventMatrix <- stringdistmatrix(EventTypes, EventTypes, method="jw")
rownames(distanceEventMatrix) <- EventTypes
EventTypesHC <- hclust(as.dist(distanceEventMatrix))
par(mar=c(2,2,1,3))
plot(as.dendrogram(EventTypesHC),horiz=T,main="Figur-1: Event Types Dendrogram")
```

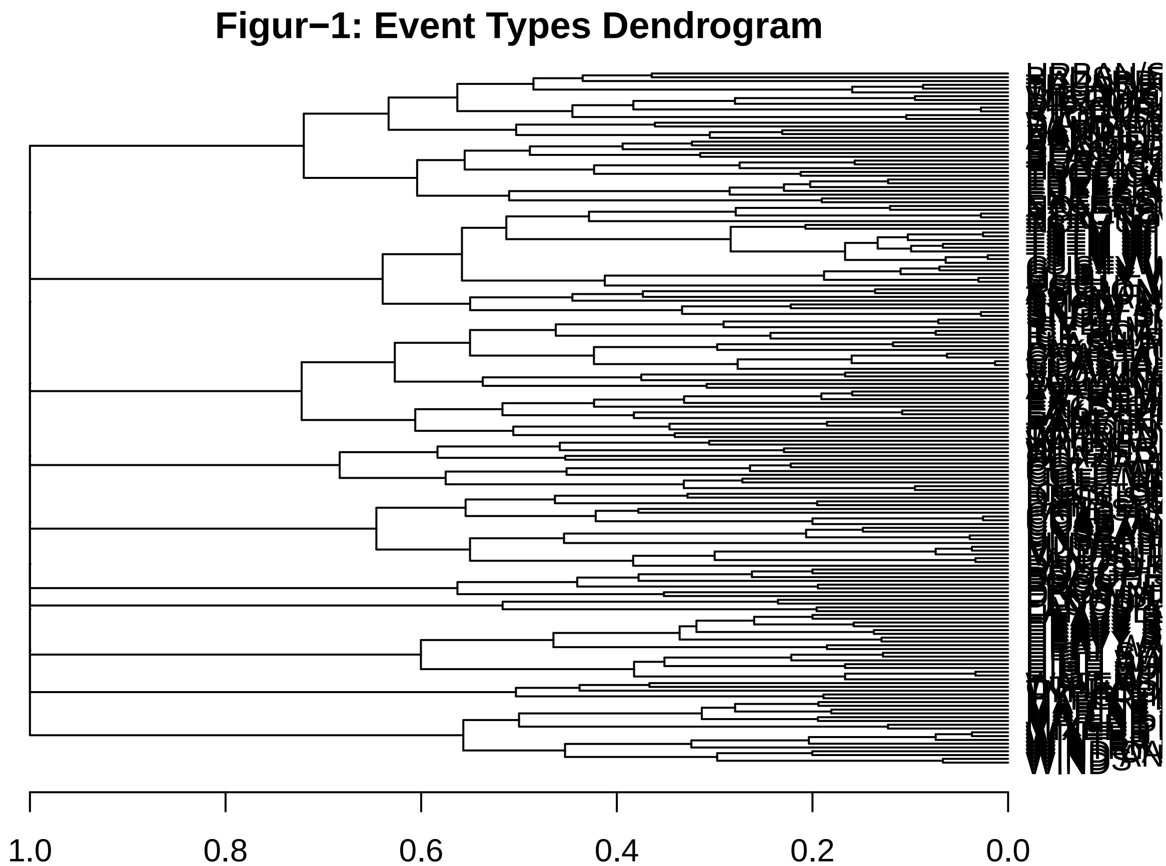

Figure-1 From this Dendrogram it is hard to point out what could be the better h value for cutree. As NOAA's National Weather Service (NWS) mentioned that we have only 48 events, we tried cutree with k=48, which produced the following sample output.

```
dend1 <- cutree(EventTypesHC,k=48)
EventTypesTable <- data.frame(EventTypes,Cluster=dend1)
TotalRows <- nrow(EventTypesTable)
rownames(EventTypesTable) <- 1:TotalRows
EventTypesTable <- arrange(EventTypesTable,Cluster)
filter(EventTypesTable,Cluster==45)
```

```
##          EventTypes Cluster
## 1 UNSEASONABLE COLD      45
## 2 UNSEASONABLY COLD      45
## 3 UNSEASONABLY WARM      45
## 4   UNSEASONAL RAIN      45
```

This sample does not make sense to combined COLD, RAIN & COLD together. We also tried with different h values and we found h=.14 is the better option for us, which produced the following sample output.

```
dend1 <- cutree(EventTypesHC,h=.14)
EventTypesTable <- data.frame(EventTypes,Cluster=dend1)
TotalRows <- nrow(EventTypesTable)
rownames(EventTypesTable) <- 1:TotalRows
```

```
EventTypesTable <- arrange(EventTypesTable,Cluster)
filter(EventTypesTable,Cluster==45)
```

```
##       EventTypes Cluster
## 1  FREEZING RAIN      45
## 2 FREEZING SPRAY      45
```

```
filter(EventTypesTable, grepl('UNSEASONA', EventTypes))
```

```
##          EventTypes Cluster
## 1 UNSEASONABLE COLD     129
## 2 UNSEASONABLY COLD     129
## 3 UNSEASONABLY WARM     130
## 4   UNSEASONAL RAIN     131
```

Merge new Culsterd Event Types With the Storm Data

After observing the above sample output we decided to accept the new EnventTypesTables which will help us to combined all the different types and get a meaningful Event Type Data set. We merged new Events types Dataset with the Storm Dataset and we used EventName column for rest of the analysis.

```
EventTypesByCluster <- EventTypesTable %>% group_by(Cluster) %>% slice(which.max(EventTypes))
colnames(EventTypesByCluster) <- c("EventName","Cluster")
Events <- full_join(EventTypesTable,EventTypesByCluster)
```

```
## Joining by: "Cluster"
```

```
stormDataset <- merge(stormDataset, Events, by.x="EVTYPE", by.y="EventTypes", all.x=T, all.y=F)
```

Property/Crop Damage Dollar Amount

As PROPDMG and CROPDMG have different unit values as follows: B-Billion, M-Million, K-Thousand. We converted all amounts fields to one dollar unit for our analysis.

```
stormDataset$PROPDMG <- ifelse(stormDataset$PROPDMGEXP == 'B', 1E9,
                        ifelse(stormDataset$PROPDMGEXP == 'M', 1E6,
                        ifelse(stormDataset$PROPDMGEXP == 'K', 1E3, 0))) * stormDataset$PROPDMG
stormDataset$CROPDMG <- ifelse(stormDataset$CROPDMGEXP == 'B', 1E9,
                        ifelse(stormDataset$CROPDMGEXP == 'M', 1E6,
                        ifelse(stormDataset$CROPDMGEXP == 'K', 1E3, 0))) * stormDataset$CROPDMG
# Summarized Dataset
stormDatasetSUM <- stormDataset %>% group_by(Cluster) %>%
  summarize(EventName=first(EventName), FATALITIES = sum(FATALITIES, na.rm=T),
            INJURIES = sum(INJURIES, na.rm=T), PROPDMG = sum(PROPDMG, na.rm=T),
            CROPDMG = sum(CROPDMG, na.rm=T)) %>% ungroup()
```

Results

Finding Top Ten Events causing highest Fatalities, Injuries, Crop and Property Damages

We searched for those Event types that caused the largest effects on population health, crop and property Damages.

```
stormDatasetFATALITIES <- transform(stormDatasetSUM, EventName = reorder(EventName,FATALITIES)) %>% arra
title1 <- "Top 10 Most Fatal by Event Type"
p1 <- ggplot(data=stormDatasetFATALITIES[1:10,]) +
        scale_x_discrete(name="Event Type") +
        scale_y_continuous(name="Total Fatalities") +
        ggtitle(title1) +
        theme_bw() +
        theme(legend.title=element_blank(), legend.position=c(.25, .65),
           legend.text = element_text(face="bold", size=6),
           legend.key.size = unit(.23, "cm"),
           legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
           axis.text.y = element_text(face="bold", size=8, angle=55),
           plot.title= element_text(face="bold", size=9),
           axis.title = element_text(face="bold", size=8),
           axis.text.x=element_blank()
        ) +
        geom_bar(stat="identity",position='dodge',aes(x=EventName, y=FATALITIES,fill=factor(EventName))

stormDatasetINJURIES <- transform(stormDatasetSUM, EventName = reorder(EventName,INJURIES)) %>% arrange
title1 <- "Top 10 Most Injuries by Event Type"
p2 <- ggplot(data=stormDatasetINJURIES[1:10,]) +
        scale_x_discrete(name="Event TYpe") +
        scale_y_continuous(name="Total Injuries") +
        ggtitle(title1) +
        theme_bw() +
        theme(legend.title=element_blank(), legend.position=c(.30, .65),
           legend.text = element_text(face="bold", size=6),
           legend.key.size = unit(.23, "cm"),
           legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
           axis.text.y = element_text(face="bold", size=8, angle=55),
           plot.title= element_text(face="bold", size=9),
           axis.title = element_text(face="bold", size=8),
           axis.text.x=element_blank()
        ) +
        geom_bar(stat="identity",position='dodge',aes(x=EventName, y=INJURIES,fill=factor(EventName)),


stormDatasetCROPDMG <- transform(stormDatasetSUM, EventName = reorder(EventName,CROPDMG)) %>% arrange(d
title1 <- "Top 10 Most Crop Damage by Event Type"
p3 <- ggplot(data=stormDatasetCROPDMG[1:10,]) +
        scale_x_discrete(name="Event Type") +
        scale_y_continuous(name="US Dollars") +
        ggtitle(title1) +
        theme_bw() +
        theme(legend.title=element_blank(), legend.position=c(.25, .65),
           legend.text = element_text(face="bold", size=6),
           legend.key.size = unit(.23, "cm"),
```

```
            legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
            axis.text.y = element_text(face="bold", size=8, angle=55),
            plot.title= element_text(face="bold", size=9),
            axis.title = element_text(face="bold", size=8),
            axis.text.x=element_blank()
        ) +
        geom_bar(stat="identity",position='dodge',aes(x=EventName, y=CROPDMG,fill=factor(EventName)), s

stormDatasetPROPDMG <- transform(stormDatasetSUM, EventName = reorder(EventName,PROPDMG)) %>% arrange(de
title1 <- "Top 10 Most Property Damage by Event Type"
p4 <- ggplot(data=stormDatasetPROPDMG[1:10,]) +
        scale_x_discrete(name="Event Type") +
        scale_y_continuous(name="US Dollars") +
        ggtitle(title1) +
        theme_bw() +
        theme(legend.title=element_blank(), legend.position=c(.25, .65),
            legend.text = element_text(face="bold", size=6),
            legend.key.size = unit(.23, "cm"),
            legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
            axis.text.y = element_text(face="bold", size=8, angle=55),
            plot.title= element_text(face="bold", size=9),
            axis.title = element_text(face="bold", size=8),
            axis.text.x=element_blank()
        ) +
        geom_bar(stat="identity",position='dodge',aes(x=EventName, y=PROPDMG,fill=factor(EventName)), s

grid.arrange(p1, p2, p3, p4, ncol=2, top ="Figure-2")
```

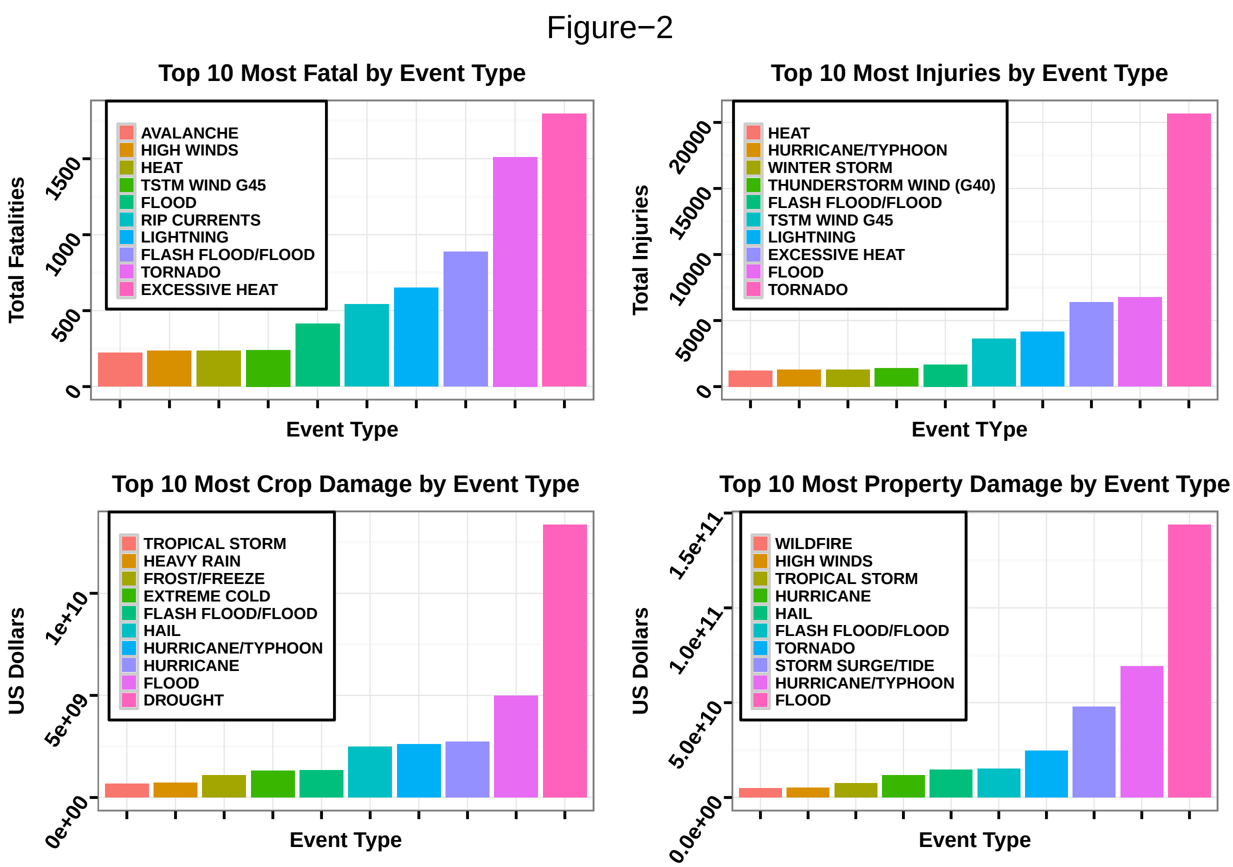

Figure-2 EXCESSIVE HEAT and TORNADO caused the highest Fatalities. TORNADO and FLOOD caused the highest Injurious. DROUGHT caused the highest Crop damages. FLOOD caused the highest Property damages.

Finding Top Ten States had highest Fatalities, Injuries, Crop and Property Damages

We searched for those Sates that have the largest effects on population health, crop and property Damage

```
# Summarized Dataset by STATEOFFICE
stormDatasetStateSUM <- stormDataset %>% group_by(STATEOFFIC) %>%
  summarize(FATALITIES = sum(FATALITIES, na.rm=T),
            INJURIES = sum(INJURIES, na.rm=T), PROPDMG = sum(PROPDMG, na.rm=T),
            CROPDMG = sum(CROPDMG, na.rm=T)) %>%
  arrange(desc(PROPDMG)) %>%
  ungroup()

stormDatasetStateFATALITIES <- transform(stormDatasetStateSUM, STATEOFFIC = reorder(STATEOFFIC,FATALITI
title1 <- "Top 10 Most Fatal by State Office"
p1 <- ggplot(data=stormDatasetStateFATALITIES[1:10,]) +
         scale_x_discrete(name="State Office") +
         scale_y_continuous(name="Total Fatalities") +
         ggtitle(title1) +
         theme_bw() +
         theme(legend.title=element_blank(), legend.position=c(.3, .7),
```

```
            legend.text = element_text(face="bold", size=5),
            legend.key.size = unit(.23, "cm"),
            legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
            axis.text.y = element_text(face="bold", size=8, angle=55),
            plot.title= element_text(face="bold", size=9),
            axis.title = element_text(face="bold", size=8),
            axis.text.x=element_blank()
         ) +
         geom_bar(stat="identity",position='dodge',aes(x=STATEOFFIC, y=FATALITIES,fill=factor(STATEOFFI

stormDatasetStateINJURIES <- transform(stormDatasetStateSUM, STATEOFFIC = reorder(STATEOFFIC,INJURIES))
title1 <- "Top 10 Most Injuries by State Office"
p2 <- ggplot(data=stormDatasetStateINJURIES[1:10,]) +
         scale_x_discrete(name="State Office") +
         scale_y_continuous(name="Total Injuries") +
         ggtitle(title1) +
         theme_bw() +
         theme(legend.title=element_blank(), legend.position=c(.35, .7),
            legend.text = element_text(face="bold", size=5),
            legend.key.size = unit(.23, "cm"),
            legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
            axis.text.y = element_text(face="bold", size=8, angle=55),
            plot.title= element_text(face="bold", size=9),
            axis.title = element_text(face="bold", size=8),
            axis.text.x=element_blank()
         ) +
         geom_bar(stat="identity",position='dodge',aes(x=STATEOFFIC, y=INJURIES,fill=factor(STATEOFFIC))


stormDatasetStateCROPDMG <- transform(stormDatasetStateSUM, STATEOFFIC = reorder(STATEOFFIC,CROPDMG)) %
title1 <- "Top 10 Most Crop Damage by State Office"
p3 <- ggplot(data=stormDatasetStateCROPDMG[1:10,]) +
         scale_x_discrete(name="State Office") +
         scale_y_continuous(name="US Dollars") +
         ggtitle(title1) +
         theme_bw() +
         theme(legend.title=element_blank(), legend.position=c(.3, .7),
            legend.text = element_text(face="bold", size=5),
            legend.key.size = unit(.23, "cm"),
            legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
            axis.text.y = element_text(face="bold", size=8, angle=55),
            plot.title= element_text(face="bold", size=9),
            axis.title = element_text(face="bold", size=8),
            axis.text.x=element_blank()
         ) +
         geom_bar(stat="identity",position='dodge',aes(x=STATEOFFIC, y=CROPDMG,fill=factor(STATEOFFIC))

stormDatasetStatePROPDMG <- transform(stormDatasetStateSUM, STATEOFFIC = reorder(STATEOFFIC,PROPDMG)) %
title1 <- "Top 10 Most Property Damage by State Office"
p4 <- ggplot(data=stormDatasetStatePROPDMG[1:10,]) +
         scale_x_discrete(name="State Office") +
         scale_y_continuous(name="US Dollars") +
         ggtitle(title1) +
```

```
        theme_bw() +
        theme(legend.title=element_blank(), legend.position=c(.25, .7),
           legend.text = element_text(face="bold", size=5),
           legend.key.size = unit(.23, "cm"),
           legend.background = element_rect(colour = "black"), legend.margin = unit(.5, "cm"),
           axis.text.y = element_text(face="bold", size=8, angle=55),
           plot.title= element_text(face="bold", size=9),
           axis.title = element_text(face="bold", size=8),
           axis.text.x=element_blank()
        ) +
        geom_bar(stat="identity",position='dodge',aes(x=STATEOFFIC, y=PROPDMG,fill=factor(STATEOFFIC))

grid.arrange(p1, p2, p3, p4, ncol=2, top ="Figure-3")
```

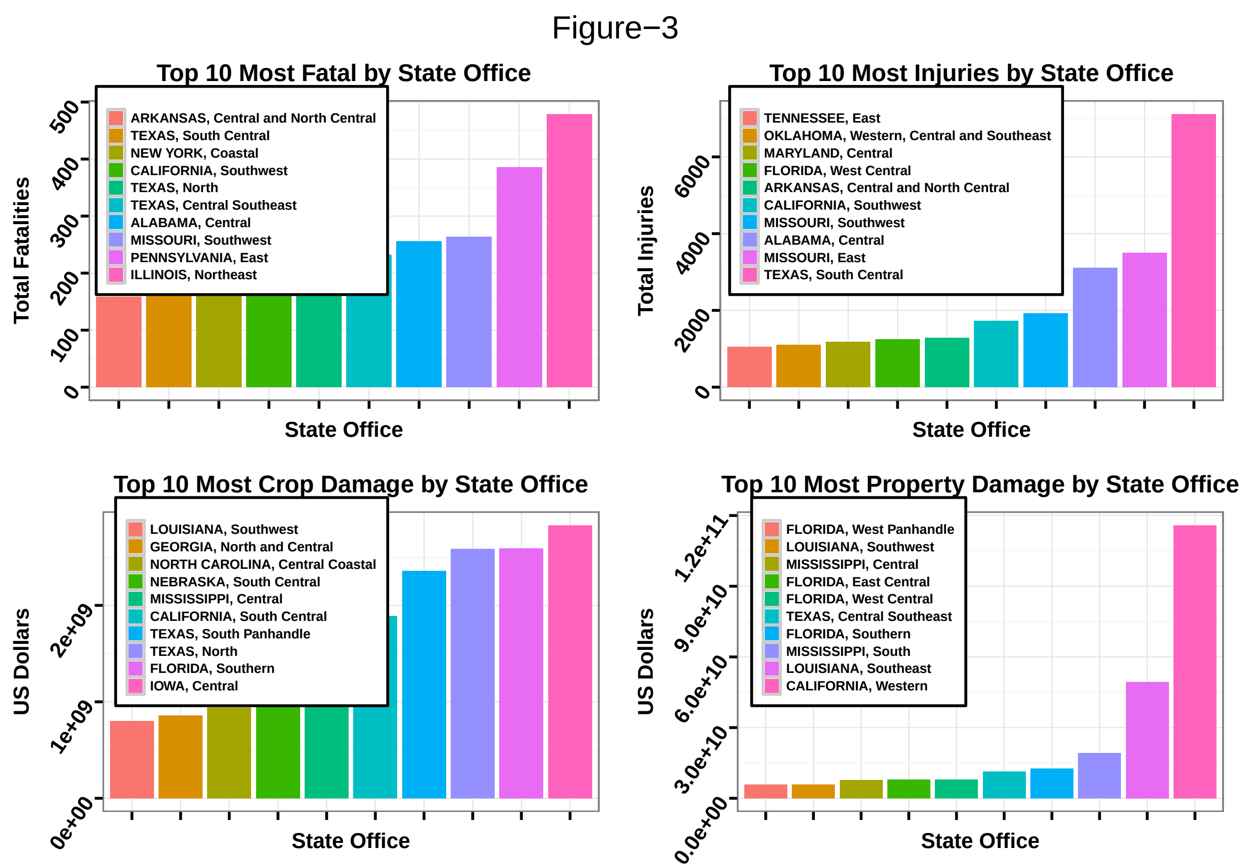

Figure-3 ILLINOIS(North), PENNSYLVANIA had the highest Fatalities. TEXAS (South), MISSOURI (East) had the highest Injurious. IOWA (Central), FLORIDA (Southarn) had the highest Crop damages. CALIFORNIA, LOUISIANA had the highest Property damages.

Fatalities, Injuries, Crop and Property Damages By Year.

We summarized all damages by Years in a Table to show the effects on population health, Crop or Property.

```r
stormDatasetYearSUM <- stormDataset %>% group_by(EventYear) %>%
  summarize(FATALITIES = sum(FATALITIES, na.rm=T),
            INJURIES = sum(INJURIES, na.rm=T), PROPDMG = sum(PROPDMG, na.rm=T),
            CROPDMG = sum(CROPDMG, na.rm=T)) %>%
  arrange(desc(EventYear)) %>%
  ungroup()

colnames(stormDatasetYearSUM) <- c("Year","Fatality","Injury","Property(US $)","Crop(US $)")
t1 <- ttheme_default(core=list(
        fg_params=list(fontface=c(rep("plain", 16), "bold.italic")),
        bg_params = list(fill=c(rep(c("grey95", "grey90"),
                                    length.out=16), "#6BAED6"),
                         alpha = rep(c(1,0.5), each=16))
        ))

grid.table(stormDatasetYearSUM, theme = t1, rows=NULL)
```

Year	Fatality	Injury	Property(US $)	Crop(US $)
2011	1002	7792	20888981960	666742000
2010	425	1855	9246487640	1785286000
2009	333	1354	5227204130	522220000
2008	488	2703	15568383080	2209793000
2007	421	2191	5788934160	1691152000
2006	599	3368	121937434190	3534238700
2005	469	1834	96789791170	4035202300
2004	370	2426	25346598870	1452177850
2003	443	2931	10254548240	1143070350
2002	498	3155	4100882450	1410368140
2001	469	2721	10027043670	1816728100
2000	477	2803	5621428050	3329170600
1999	908	5148	8721226550	3532284100
1998	687	11177	11603795630	4507685350
1997	601	3800	9558060240	1228079400
1996	542	2717	6086815350	1888530840

Table-1 Year 2011 and 1999 had the highest Fatalities. Year 2011 and 1998 had the highest Injuries. Year 2005 and 1998 had the highest Crop damages. Year 2005 and 2006 had the highest Property damages.