severe_storms.Rmd

Giuseppa Cefalu

October 11, 2019

```
knitr::opts_chunk$set(fig.width=10, fig.height=10)
```

```
##
## 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

##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine

##
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':
##
##     last_plot

## The following object is masked from 'package:stats':
##
##     filter

## The following object is masked from 'package:graphics':
##
##     layout

## Storing 'username' as the environment variable 'plotly_username'

## Storing 'key' as the environment variable 'plotly_api_key'

## Warning: 'plotly' is deprecated.
## Use 'ggplotly' instead.
## See help("Deprecated")

## Warning: 'plotly' is deprecated.
## Use 'plot_ly' instead.
## See help("Deprecated")
```

```
rm(list = ls())
graphics.off()
```

"'

```
library(dplyr)
```

```r
library(ggplot2)

library(gridExtra)

library(knitr)

library(tinytex)

library(plotly)
```

Author: Giuseppa Cefalu - giuseppacefalu@gmail.com

HEALTH AND ECONOMIC IMPACT CAUSED BY SEVERE STORMS ACROSS THE UNITED STATES

SYNOPSIS.- This document is an analysis of the health and economic impact caused by severe storms across the United States. The analysis is based on the number of personal injuries, property damage and crop damage caused by a specific weather event in a particular state in the United States. The variables included in the analysis are "INJURIES", "PROPDMG" and "CROPDMG". The data file for the analysis was obtained from 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.

In order to plot the number of injnuries, the property damage, and the crop damage versus storm event type, the data are summarized in the following way: for each of the 3 variables, first the values for each record for the corresponding event type are added in each state to obtain a single value per event type. Afterwards, the events with the highest value for the measured variable in each state are selected. These values are ploted against event type. Tables of summary statistics are added to the plot.

The study shows that the event types that cause the greates amount of health and economic impact are tornados, flush floods,tropical storms and hail.

DATA PROCESSING - The path to the data file is defined and stored in the variable data_path and the data from the file are read and stored as a dataframe in the variable name "df". The file contains disguised duplicates; for exemple, the same event type may be some times written in lower case and some other times in upper case. In order to avoid considering these duplicates as different events when their values are added, the data in the file (event type) are all converted to lower case.

Additionally, the records with v,alue of zero for the corresponding variable (INJURIES,PROPDMG and CROPDMG)are eliminated in order to decrease the number of records to be ploted since zero values are not esential for additions and for the capturing of maximum values. Furthermore, due to the large amount of records for the varible property damage, the data frame storing these data is divided into subdataframes filtering by a group of states. As a result, 4 data frames will be ploted for the category PROPDMG versus event type.Summary statistics for each data frame set are calulated and displayed in the plot. A histogram of each measures variable does not describe a normal distribution, therefore, I tried to normalize the data using Tukey-transform, however the procedure did not make the distribution more normal. It is obvious that there are extreme values that could be interpreted as outliers, but they are not outliers, they represent the maxima or the highest value of the measured variable of a Poisson distributionas that models count data on an interval of time and therefore the maxima falls within the Poisson distribution. Even though the data see ms to describe a Poisson distribution, I used the normal distribution mean and standard deviation. The parameters for a Poisson distribution could be obtained in R using the function MASS::fitdistr(mean(data),"Poisson")

Define path to file

```
data_path <- file.path("~", "DataScience","Data Files","STORM DATA",
"repdata_data_StormData.csv.bz2")

# Read data file
data <- read.csv(data_path, stringsAsFactors = FALSE)

# Convert to lower case to process duplicates as single items(TORNADO/tornado)
data$EVTYPE <- tolower(data$EVTYPE)

# Eliminate unnecessary values for records containing injuries
df <- filter(data, INJURIES > 0)

# This counts were done to plot a histogram of the distribution,
# however the plot is not included here.
h <- df %>% group_by(EVTYPE) %>% summarize(frequency = n())

# Types of events more harmful to the population. Here the
#sum of the values of the variable INJURIES for each event
```

```r
#type and state are added.
df <- df %>% group_by(STATE, EVTYPE) %>% summarize(INJURIES =
sum(INJURIES, na.rm = TRUE))

# Here the events causing the largest number of injuries in each state is extracted
df_max <- df %>% group_by(EVTYPE) %>% summarize(INJURIES =
max(INJURIES))

# Summary statistics for df_max
stats <- df_max %>% summarize(MEAN = mean(INJURIES),
MEDIAN = median(INJURIES), MAX = max(INJURIES), MIN =
min(INJURIES),
SDEV = sd(INJURIES))

# Eliminate unnecessary values for records containing property values
df <- filter(data, PROPDMG > 0)

# Types of events that have greater economic consequences
# (property damage).  Calculate the sum of values for each
# event type and for each state across the United States.
cost1 <- df %>% group_by(STATE, EVTYPE) %>% summarize(PROPDMG =
sum(PROPDMG, na.rm = TRUE))

# Select the states of GU, TX, AK, TN, CA, KY, VI, IL, WI,
# MS, UT, CO, ND.
cost1a <- filter(cost1, STATE == "GU" | STATE == "TX" | STATE == "AK"
| STATE == "TN" | STATE == "CA" | STATE == "KY" | STATE == "VI" |
STATE == "IL" | STATE == "WI" | STATE == "MS" | STATE == "UT" |
STATE == "CO" | STATE == "ND" )

# Types of events that have the greatest economic consequences
# (property damage). Calculate the maximum value among all
# the event types values in each state a cross the United States
cost1a_max <- cost1a %>% group_by(EVTYPE) %>% summarize(PROPDMG = max(PROPDMG))

# Select the states of SD, FL, IA, PR, MA, DC, KS, AR, GA, OR,
# AZ, DE, and GM
cost1b <- filter(cost1, STATE == "SD" | STATE == "FL" | STATE == "IA"
| STATE == "PR" | STATE == "MA" | STATE == "DC" | STATE == "KS"
| STATE == "AR" | STATE == "GA" | STATE == "OR" | STATE == "AZ"
| STATE == "DE" | STATE == "GM")

# Types of events having the greatest economic consequences
# (property damage). Extract the event with maximum value for
# each of the measured variable in each state
cost1b_max <- cost1b %>% group_by(EVTYPE) %>% summarize(PROPDMG = max(PROPDMG))

# Select the states of WA, NC, NH, NJ, NY, NE, ID, HI, IN, NV,
# AS, OH, LA, and VT
cost1c <- filter(cost1, STATE == "WA" | STATE == "NC" | STATE == "NH"
| STATE == "NJ" | STATE == "NY" | STATE == "NE" | STATE == "ID"
| STATE == "HI" | STATE == "IN" | STATE == "NV" | STATE == "AS"
| STATE == "OH" | STATE == "LA" | STATE == "VT" )
```

```r
# Types of Events Having the Greatest Economic Consequences
# (property damage). Extract the events with the maximum value
# for the measured variable in each state

cost1c_max <- cost1c %>% group_by(EVTYPE) %>% summarize(PROPDMG = max(PROPDMG))

# Select the states of NM, WY, MD, OK, MI, MT, WV, ME, PA, UT,
# SC, CT, LS, LE, NZ, PZ, RI, and VA
cost1d <- filter(cost1, STATE == "NM" | STATE == "WY" | STATE == "MD"
| STATE == "OK" | STATE == "MI" | STATE == "MT" | STATE == "WV" |
STATE == "ME" | STATE == "PA" | STATE == "UT"| STATE == "SC" |
STATE == "CT" | STATE == "LS" | STATE == "LE" | STATE == "NC" |
STATE == "PZ" | STATE == "RI" | STATE == "VA")

# Types of Events Having the Greatest Economic Consequences
# (property damage). Extract the events with the maximum value
# of the measured variable in each state
cost1d_max <- cost1d %>% group_by(EVTYPE) %>% summarize(PROPDMG = max(PROPDMG))

# Calculate summary statistics for the cost1a_max set
statsa <- cost1a_max %>% summarize(MEAN = mean(PROPDMG),
MEDIAN = median(PROPDMG), MAX = max(PROPDMG), MIN = min(PROPDMG),
SDEV = sd(PROPDMG))

# Calculate summary statics for the cost1b_max set
statsb <- cost1b_max %>% summarize(MEAN = mean(PROPDMG),
MEDIAN = median(PROPDMG), MAX = max(PROPDMG), MIN = min(PROPDMG),
SDEV = sd(PROPDMG))

# Calculate summary stats for cost1b_max set
statsd <- cost1d_max %>% summarize(MEAN = mean(PROPDMG),
MEDIAN = median(PROPDMG), MAX = max(PROPDMG), MIN = min(PROPDMG),
SDEV =sd(PROPDMG))

# Calculate summary stats for the cost1c_max set
statsc <- cost1c_max %>% summarize(MEAN = mean(PROPDMG),
MEDIAN = median(PROPDMG), MAX = max(PROPDMG), MIN = min(PROPDMG),
SDEV = sd(PROPDMG))

# Select data containing the crop damage across the USA
# and eliminate unnececsary records. Add the values of
# crop damage for each state and for each event type
cost2 <- df %>% group_by(STATE, EVTYPE) %>% summarize(CROPDMG =
sum(CROPDMG, na.rm = TRUE))

# Select the maximum values of crop damage across all the states in the USA
cost2_max <- cost2 %>% group_by(EVTYPE) %>% summarize(CROPDMG =
max(CROPDMG))

# Add the maximum values of crop damage for each state and for each event type
cost2 <- df %>% group_by(STATE, EVTYPE) %>% summarize(CROPDMG =
sum(CROPDMG, na.rm = TRUE))
```

```r
# Select the maximum values of crop damage across all the states in the USA
cost2_max <- cost2 %>% group_by(EVTYPE) %>% summarize(CROPDMG =
max(CROPDMG))

# Calculate summary stats for cost2 set
stats2 <- cost2_max %>% summarize(MEAN = mean(CROPDMG),
MEDIAN = median(CROPDMG), MAX = max(CROPDMG), MIN = min(CROPDMG),
SDEV = sd(CROPDMG))
```

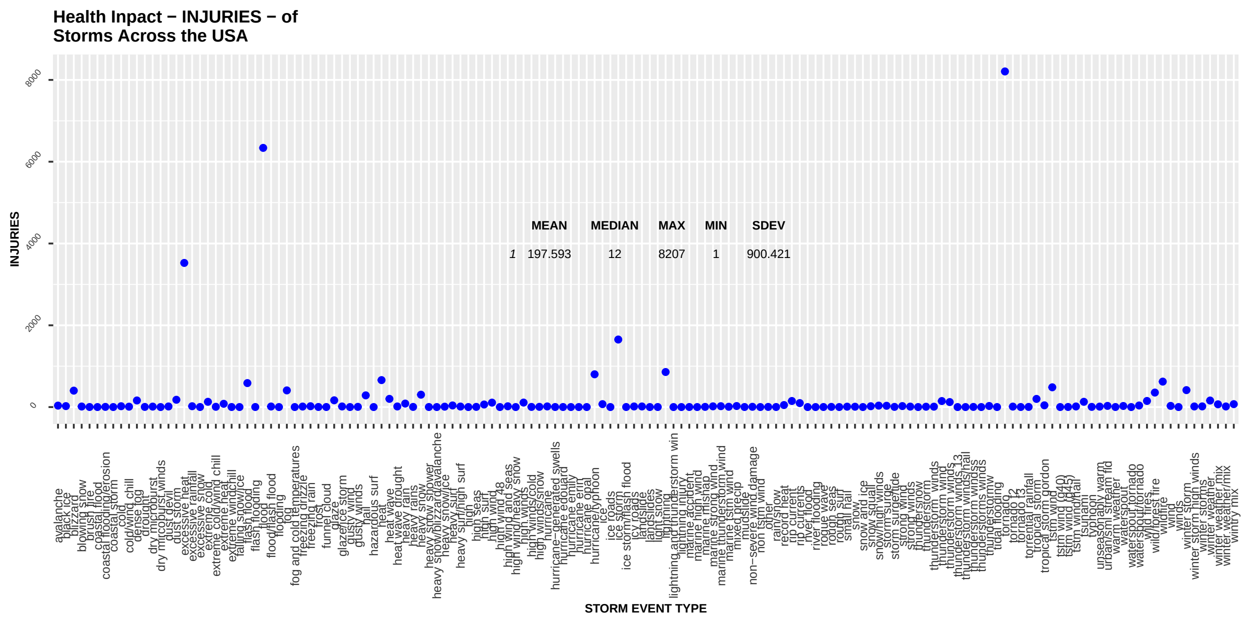
RESULTS

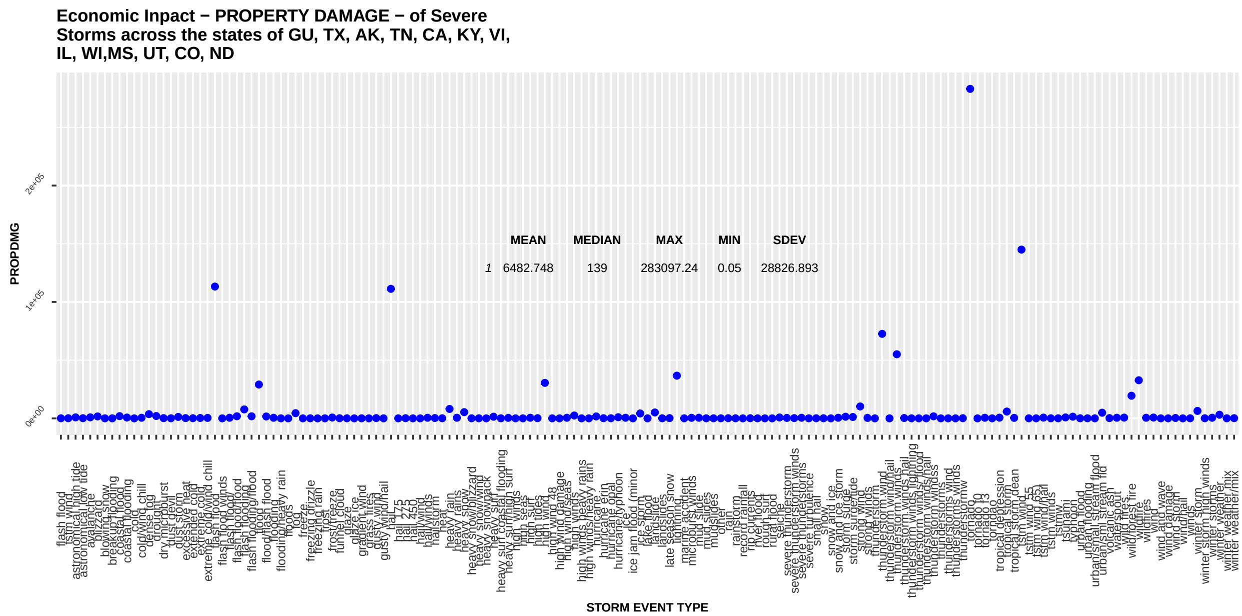
Initialize device

```r
dev.new(width =20000, height = 10000, unit = "px")
```

```r
library(plotly)
# Plot the event types that cause the largest number of injuries across
# all the states in the USA. Include the summary statistics table
# on the graph
p1 <- ggplot(df_max, aes(EVTYPE, INJURIES)) + geom_point(color = "blue") +
theme(legend.position="top") +
theme(axis.text.x = element_text(color = "grey20", size = 7.0,
angle = 90, hjust = .5, vjust = .5, face = "plain"),
axis.text.y = element_text(color = "grey20", size = 5,
angle = 50,
hjust = 0.5, vjust = 0, face = "plain")) +
ggtitle("Health Inpact - INJURIES - of
Storms Across the USA") + theme(plot.title = element_text(size = 10,
face = "bold")) +
xlab("STORM EVENT TYPE") +
theme(axis.title.y = element_text(size = rel(0.5), angle = 90),
axis.title=element_text(size=14,face="bold")) +
theme(axis.title.x = element_text(size = rel(0.5), angle = 0),
axis.title=element_text(size=14,face="bold")) +
annotation_custom(tableGrob(round(stats, digits = 3),
theme=ttheme_minimal(base_size = 7)))

# Plot the events causing the maximum amount of property damage
#across cost1a states. Include the summary statistics table on the graph.
p2 <- ggplot(cost1a_max, aes(EVTYPE, PROPDMG)) +
geom_point(color = "blue") +
theme(legend.position="top") +
theme(axis.text.x = element_text(color = "grey20",
size = 7.0, angle = 90, hjust = .5,
vjust = .5, face = "plain"),
axis.text.y = element_text(color = "grey20",
size = 5, angle = 50, hjust = 0.5, vjust = 0,
face = "plain")) +
ggtitle("Economic Inpact - PROPERTY DAMAGE - of Severe
Storms across the states of GU, TX, AK, TN, CA, KY, VI,
IL, WI,MS, UT, CO, ND") +
theme(plot.title = element_text(size = 10,
```

```r
face = "bold"), axis.title=element_text(size=14,face="bold")) +
xlab("STORM EVENT TYPE") +
theme(axis.title.y = element_text(size = rel(0.5),
angle = 90), axis.title=element_text(size = 14,
face = "bold")) +
theme(axis.title.x = element_text(size = rel(0.5),
angle = 0), axis.title=element_text(size=14,face="bold")) +
annotation_custom(tableGrob(round(statsa, digits = 3), theme=ttheme_minimal(base_size = 7)))


# Plot 2 graphs on the same panel
grid.arrange(p1, p2)
```





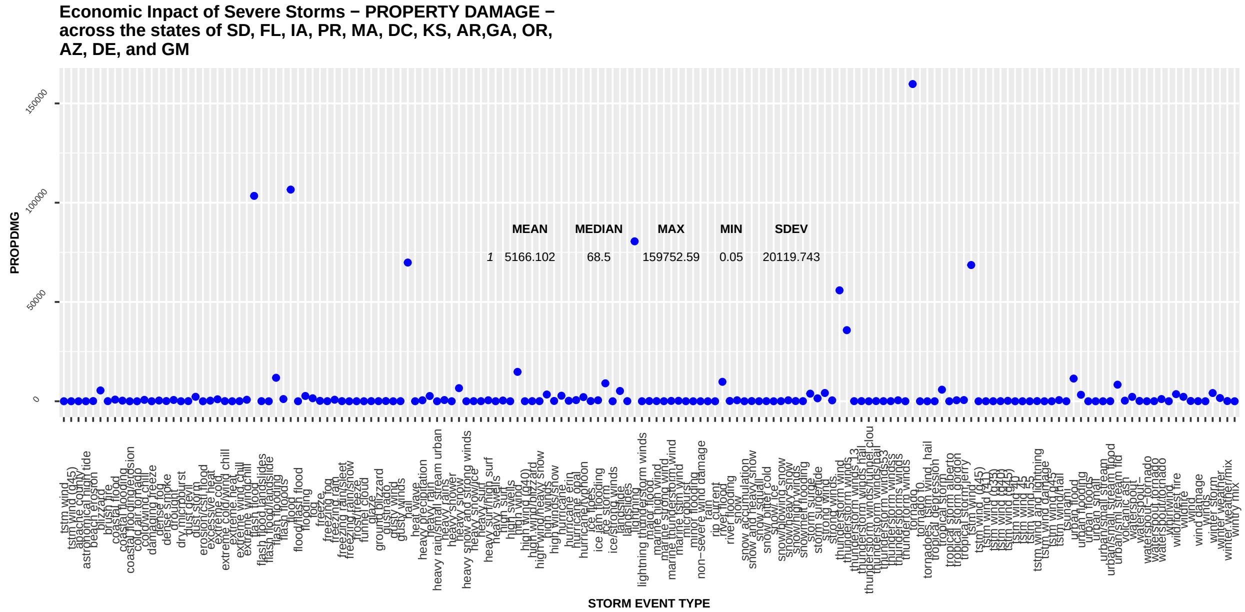
As can be seen in the plots above, the graetest health impact to humans across the united states is caused by tornados, and flash flooding, with a large difference between the mean and the maximum score, and the greater economic impact with respect to property damage in the States of GU, TX, AK, TN, CA, KY, VI, IL, WI, MS, UT, CO, ND is caused by tornados and tropical storms, with a large difference between the mean and the maximum score.

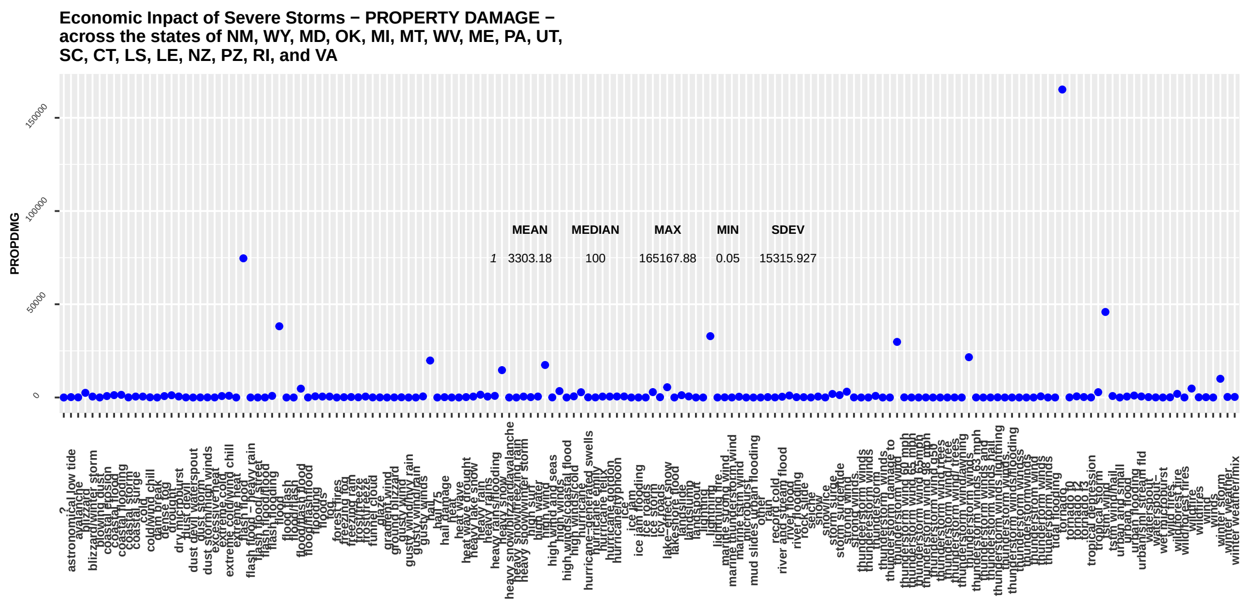
Initialize device

dev.new(width =20000, height = 10000, unit = "px")

```
library(plotly)
# Plot the events causing the Largest ammount of property
#damage across cost1b states. Include the summary statistics
#table on the graph
p3 <- ggplot(cost1b_max, aes(EVTYPE, PROPDMG)) +
geom_point(color = "blue") +
theme(legend.position="top") +
theme(axis.text.x = element_text(color = "grey20",
size = 7.0, angle = 90, hjust = .5, vjust = .5,
face = "plain"),
axis.text.y = element_text(color = "grey20",
size = 5, angle = 50, hjust = 0.5, vjust = 0,
face = "plain")) +
ggtitle("Economic Inpact of Severe Storms - PROPERTY DAMAGE -
across the states of SD, FL, IA, PR, MA, DC, KS, AR,GA, OR,
AZ, DE, and GM") +
theme(plot.title = element_text(size = 10,
face = "bold")) +
xlab("STORM EVENT TYPE") +
theme(axis.title.y = element_text(size = rel(0.5),
angle = 90), axis.title=element_text(size=14,
face="bold")) +
theme(axis.title.x = element_text(size = rel(0.5),
angle = 0), axis.title=element_text(size=14,
face="bold")) +
annotation_custom(tableGrob(round(statsb, digits = 3), theme=ttheme_minimal(base_size = 7)))

# Select data containing the crop damage across the USA and eliminate unnececsary records.
df <- filter(data, CROPDMG > 0)

# Plot the events causing the maximum amount of property
#damage across cost1d states. Include the summary statistics
#table on the plot
p4 <- ggplot(cost1d_max, aes(EVTYPE, PROPDMG)) +
geom_point(color = "blue") + theme(legend.position="top") +
 theme(axis.text.x = element_text(color = "grey20",
siz = 7.0, angle = 90, hjust = .5, vjust = .5, face = "bold"),
axis.text.y = element_text(color = "grey20", size = 5,
angle = 50, hjust = 0.5, vjust = 0, face = "plain")) +
```

```r
ggtitle("Economic Inpact of Severe Storms - PROPERTY DAMAGE -
across the states of NM, WY, MD, OK, MI, MT, WV, ME, PA, UT,
SC, CT, LS, LE, NZ, PZ, RI, and VA") +
theme(plot.title = element_text(size = 10, face = "bold")) +
xlab("STORM EVENT TYPE") +
theme(axis.title.y = element_text(size = rel(0.5),
angle = 90),
axis.title=element_text(size=14,face="bold")) +
theme(axis.title.x = element_text(size = rel(0.5),
angle = 0),
axis.title=element_text(size=14,face="bold")) +
annotation_custom(tableGrob(round(statsd, digits = 3), theme=ttheme_minimal(base_size = 7)))

# Plot 2 graphs on the same panel
grid.arrange(p3, p4)
```





The above plots show that the greater economic impact with respect to property damage in the states of SD, FL, IA, PR, MA, DC, KS, AR, GA, OR, AZ, DE, and GM is due to tornados, floods and flash floods, with a large difference between the mean and the maximum score, and the greates economic impact with respect to property damage in the states of NM, WY, MD, OK, MI, MT, WV, ME, PA, UT, SC, CT, LS, LE, NZ, PZ, RI, and VA is caused by tidal flooding with a large difference between the mean and the maximum socre.

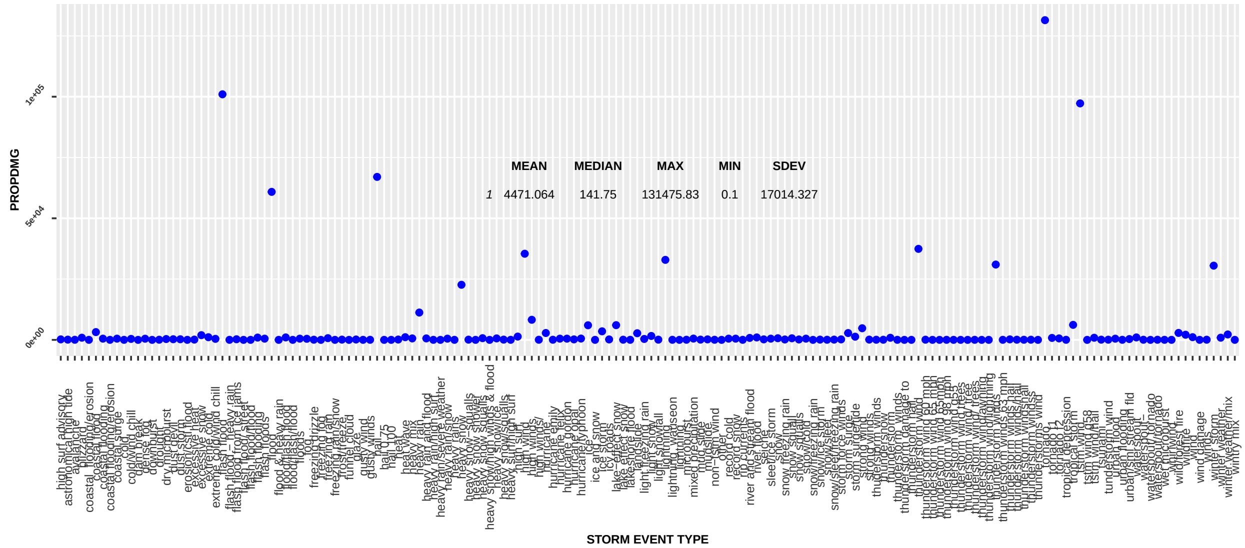
Initialize device

dev.new(width =20000, height = 10000, unit = "px")

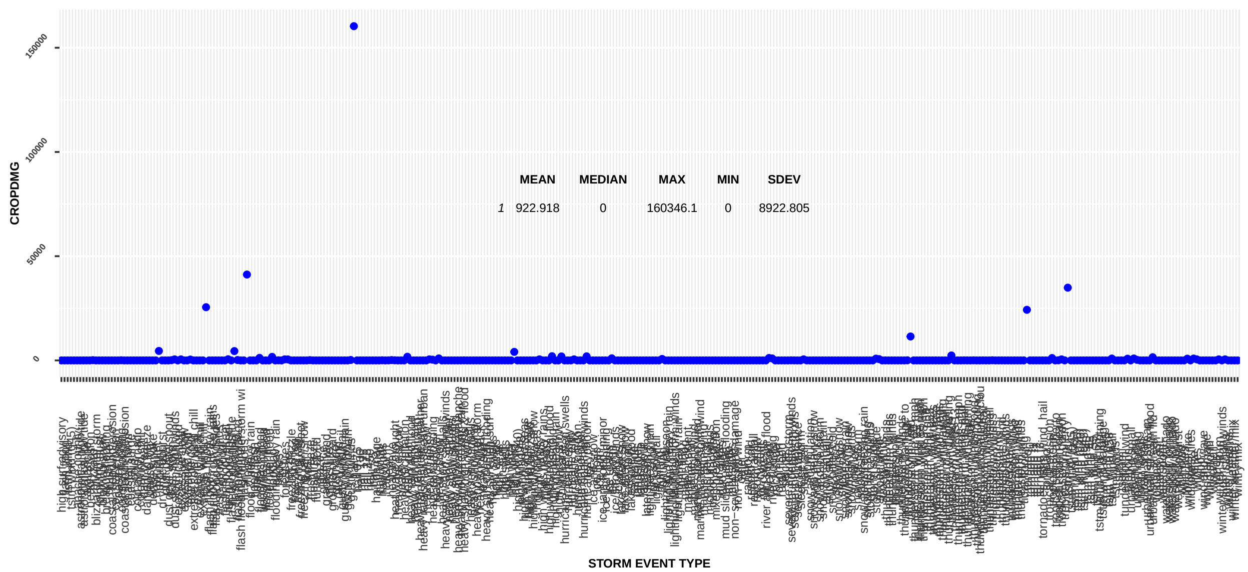
```
library(plotly)
# Plot the events causing the gratest amount of property
#damage across cost1c states. Include the summary statistics
```

```r
#table on the graph
p5 <- ggplot(cost1c_max, aes(factor(EVTYPE), PROPDMG)) +
geom_point(color = "blue") + theme(legend.position="top") +
theme(axis.text.x = element_text(color = "grey20", size = 7.0,
angle = 90, hjust = .5, vjust = .5, face = "plain"),
axis.text.y = element_text(color = "grey20", size = 5,
angle = 50, hjust = 0.5, vjust = 0, face = "bold")) +
ggtitle("Economic Inpact of Severe Storms - PROPERTY DAMAGE -
across the states of WA, NC, NH, NJ, NY, NE, ID, HI,IN, NV, AS,
OH, LA, and VT") +
theme(plot.title = element_text(size = 10, face = "bold")) +
xlab("STORM EVENT TYPE") +
theme(axis.title.y = element_text(size = rel(0.5),
angle = 90), axis.title=element_text(size=14,face="bold")) +
theme(axis.title.x = element_text(size = rel(0.5),
angle = 0), axis.title=element_text(size=14,face="bold")) +
annotation_custom(tableGrob(round(statsc, digits = 3), theme=ttheme_minimal(base_size = 7)))

# Plot events causing the largest amount of crop
#damage across the USA. Include the summary statistics
#table on the graph.
p6 <- ggplot(cost2_max, aes(EVTYPE, CROPDMG)) +
geom_point(color = "blue") + theme(legend.position="top") +     theme(axis.text.x = element_text(color =
angle = 90, hjust = .5, vjust = .5, face = "plain"),
axis.text.y = element_text(color = "grey20", size = 5,
angle = 50, hjust = 0.5, vjust = 0, face = "bold")) +
ggtitle("Economic Impact of Severe Storms - CROP DAMAGE -
across the united states") +
theme(plot.title = element_text(size = 10, face = "bold")) +
xlab("STORM EVENT TYPE") +
theme(axis.title.y = element_text(size = rel(0.5),
angle = 90), axis.title=element_text(size=14,face="bold")) +
theme(axis.title.x = element_text(size = rel(0.5),
angle = 0), axis.title=element_text(size=14,face="bold")) +
annotation_custom(tableGrob(round(stats2, digits = 3), theme=ttheme_minimal(base_size = 7)))

# Plot 2 graphs on the same panel
grid.arrange(p5,p6)
```

Economic Impact of Severe Storms − PROPERTY DAMAGE − across the states of WA, NC, NH, NJ, NY, NE, ID, HI,IN, NV, AS, OH, LA, and VT



Economic Impact of Severe Storms − CROP DAMAGE − across the united states



The last two plots show that in the states of WA, NC, NH, NJ, NY, NE, ID, HI,IN, NV, AS, OH, LA, and VT the gratest economic impact with respect to property damage is caused by tornados flash floods and tropical storm with a large difference between the mean and the maximum score, and the gratest economic impact with respect to crop damage across the United Stats is caused by hail with a large difference between the mean and the maximum value.

Across the Unitad states the gratest helth and economic ipact is caused by torbados, flash flows, tropical storms and hail.