Tufte Handout

John Smith

August 13th, 2014

Synopsis

The U.S. National Oceanic and Atmospheric Administration's (NOAA) storm database[1] 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. The events in the database correspond to 902,297 observations starting in the year 1950 and ending in November 2011.

[1] http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2

This report addresses the questions:

1. Across the United States, which types of events are most harmful with respect to population health?

2. Across the United States, which types of events have the greatest economic consequences?

This report was created for a Coursera class[2] on the following environment:

[2] http://class.coursera.org/repdata-008

	_
platform	i386-w64-mingw32
arch	i386
os	mingw32
system	i386, mingw32
status	
major	3
minor	1.1
year	2014
month	07
day	10
svn rev	66115
language	R
version.string	R version 3.1.1 (2014-07-10)
nickname	Sock it to Me

Downloading the directive

1. Download the NOAA study's documentation in PDF form[3]. This file contains information about the study and the variable descriptions in the data.
2. Download the NOAA study's frequently asked questions guide in PDF form[4]. This file contains codes, abbreviations, and notes regarding the database.

[3] http://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf

[4] http://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2FNCDC%20Storm%20Events-FAQ%20Page.pdf

```
docURL <- "http://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2Fpd01016005curr.pdf"
if (!file.exists("doc.pdf")) {
    os <- (Sys.info()[["sysname"]])
    if (os == "Windows") {
        download.file(docURL, destfile = "doc.pdf",
            quiet = T, mode = "wb")
    } else {
        download.file(docURL, destfile = "doc.pdf",
            method = "curl", quiet = T, mode = "wb")
    }
}
faqURL <- "http://d396qusza40orc.cloudfront.net/repdata%2Fpeer2_doc%2FNCDC%20Storm%20Events-FAQ%20Page.pdf
if (!file.exists("faq.pdf")) {
    os <- (Sys.info()[["sysname"]])
    if (os == "Windows") {
        download.file(faqURL, destfile = "faq.pdf",
            quiet = T, mode = "wb")
    } else {
        download.file(faqURL, destfile = "faq.pdf",
            method = "curl", quiet = T, mode = "wb")
    }
}
```

Data Processing

1. Download the NOAA database in raw and compressed (.bz2) form. This will check the operating system type and if the OS is not Windows, then it uses method='curl'.
2. Read the bz2 file as csv into an R dataframe.

```
fileURL <- "http://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStormData.csv.bz2"
dateDownloaded <- date()
if (!file.exists("data")) {
    dir.create("data")
}
if (!file.exists("data/StormData.csv.bz2")) {
    if (os == "Windows") {
        download.file(fileURL, destfile = "data/StormData.csv.bz2",
            quiet = T)
    } else {
        download.file(fileURL, destfile = "data/StormData.csv.bz2",
            method = "curl", quiet = T)
    }
}
if (!suppressWarnings(require("R.utils"))) {
```

```
    install.packages("R.utils")
    library(R.utils)
}
```

```
## Loading required package: R.utils
## Loading required package: R.oo
## Loading required package: R.methodsS3
## R.methodsS3 v1.6.1 (2014-01-04) successfully loaded. See ?R.methodsS3 for help.
## R.oo v1.18.0 (2014-02-22) successfully loaded. See ?R.oo for help.
##
## Attaching package: 'R.oo'
##
## The following objects are masked from 'package:methods':
##
##     getClasses, getMethods
##
## The following objects are masked from 'package:base':
##
##     attach, detach, gc, load, save
##
## R.utils v1.34.0 (2014-10-07) successfully loaded. See ?R.utils for help.
##
## Attaching package: 'R.utils'
##
## The following object is masked from 'package:utils':
##
##     timestamp
##
## The following objects are masked from 'package:base':
##
##     cat, commandArgs, getOption,
##     inherits, isOpen, parse, warnings
```

```
if (!exists("data")) {
    data <- read.csv(bzfile("data/StormData.csv.bz2"),
        header = TRUE, stringsAsFactors = FALSE)
}
```

Analysis

Which types of events are most harmful with respect to population health? How many unique event types are there in the data?

```
numeventtypes <- length(table(data$EVTYPE))
print(numeventtypes, include.rownames = FALSE)
```

[1] 985

There are 985 unique event types in the data, and many of these event types are duplications with different unique names. Additionally, it was provided by Dr. Peng that:

"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."

Because of this, we decided to look at the number of observations by year and select only those observations that fall outside the top 80% number of observations by **BGN_DATE**. To do this, we convert the **BGN_DATE** field from a factor to a date and crate a Pareto chart of the **BGN_DATE** field[5]:

[5] http://cran.r-project.org/web/packages/qcc/index.html

```
if (!suppressWarnings(require("qcc"))) {
    install.packages("qcc")
    library(qcc)
}
```

```
NA Loading required package: qcc
NA Package 'qcc', version 2.6
NA Type 'citation("qcc")' for citing this R package in publications.
```

```
year <- as.numeric(format(as.Date(data$BGN_DATE,
    format = "%m/%d/%Y"), "%Y"))
if (!suppressWarnings(require("xtable"))) {
    install.packages("xtable")
    library(xtable)
}
```

```
NA Loading required package: xtable
```

```
options(xtable.comment = FALSE)
options(xtable.booktabs = TRUE)
xtbl_year <- xtable(pareto.chart(table(year),
    plot = FALSE), caption = "Distribution of Observations by Year")
print(xtbl_year, floating = FALSE)
```

	Frequency	Cum.Freq.	Percentage	Cum.Percent.
2011	62174.00	62174.00	6.89	6.89
2008	55663.00	117837.00	6.17	13.06
2010	48161.00	165998.00	5.34	18.40
2009	45817.00	211815.00	5.08	23.48
2006	44034.00	255849.00	4.88	28.36
2007	43289.00	299138.00	4.80	33.15
2003	39752.00	338890.00	4.41	37.56
2004	39363.00	378253.00	4.36	41.92
2005	39184.00	417437.00	4.34	46.26
1998	38128.00	455565.00	4.23	50.49
2002	36293.00	491858.00	4.02	54.51
2001	34962.00	526820.00	3.87	58.39
2000	34471.00	561291.00	3.82	62.21
1996	32270.00	593561.00	3.58	65.78
1999	31289.00	624850.00	3.47	69.25
1997	28680.00	653530.00	3.18	72.43
1995	27970.00	681500.00	3.10	75.53
1994	20631.00	702131.00	2.29	77.82
1992	13534.00	715665.00	1.50	79.32
1993	12607.00	728272.00	1.40	80.71
1991	12522.00	740794.00	1.39	82.10
1990	10946.00	751740.00	1.21	83.31
1989	10410.00	762150.00	1.15	84.47
1986	8726.00	770876.00	0.97	85.43
1983	8322.00	779198.00	0.92	86.36
1985	7979.00	787177.00	0.88	87.24
1987	7367.00	794544.00	0.82	88.06
1984	7335.00	801879.00	0.81	88.87
1988	7257.00	809136.00	0.80	89.68
1982	7132.00	816268.00	0.79	90.47
1980	6146.00	822414.00	0.68	91.15
1974	5386.00	827800.00	0.60	91.74
1975	4975.00	832775.00	0.55	92.29
1981	4517.00	837292.00	0.50	92.80
1973	4463.00	841755.00	0.49	93.29
1979	4279.00	846034.00	0.47	93.76
1976	3768.00	849802.00	0.42	94.18
1977	3728.00	853530.00	0.41	94.60
1978	3657.00	857187.00	0.41	95.00
1971	3471.00	860658.00	0.38	95.39
1968	3312.00	863970.00	0.37	95.75
1970	3215.00	867185.00	0.36	96.11
1969	2926.00	870111.00	0.32	96.43
1965	2855.00	872966.00	0.32	96.75
1967	2688.00	875654.00	0.30	97.05
1962	2389.00	878043.00	0.26	97.31
1966	2388.00	880431.00	0.26	97.58
1964	2348.00	882779.00	0.26	97.84
1961	2246.00	885025.00	0.25	98.09
1958	2213.00	887238.00	0.25	98.33

It looks like 80% of the observations fall between the years 1992 and 2011. Therefore, we decided to limit the analysis to the years 1992 through 2011. As such, we select all rows where the **BGN_DATE** is between 1992 and 2011. Additionally, we only care about the date of the event, event type, number of fatalities and injuries, and amount of property and crop damage. As such, we will limit the selected fields to the following:

- BGN_DATE
- EVTYPE
- FATALITIES
- INJURIES
- PROPDMG
- PROPDMGEXP
- CROPDMG
- CROPDMGEXP

```
data <- subset(data, BGN_DATE > as.Date("12/31/1991",
    format = "%m/%d/%Y"))
```

```
## Warning: Incompatible methods ("Ops.factor",
## "Ops.Date") for ">"
```

```
cols <- c("BGN_DATE", "EVTYPE", "FATALITIES",
    "INJURIES", "PROPDMG", "PROPDMGEXP", "CROPDMG",
    "CROPDMGEXP")
data <- data[cols]
```

Now we are ready to look at the events that caused the most fatalities between 1992 and 2011:

```
if (!suppressWarnings(require("dplyr"))) {
    install.packages("dplyr")
    library(dplyr)
}
```

```
Loading required package: dplyr

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
```

```
topfatals <- data %.% group_by(EVTYPE) %.% summarize(n = sum(FATALITIES)) %.%
    mutate(proportion = n/sum(n)) %.% arrange(desc(proportion))
top10fatals <- as.data.frame(head(topfatals, 10))
options(xtable.comment = FALSE)
options(xtable.booktabs = TRUE)
xtbl_fatals <- xtable(top10fatals, caption = "Top 10 Events Causing Fatalities")
print(xtbl_fatals, floating = FALSE, include.rownames = FALSE)
```

EVTYPE	n	proportion
EXCESSIVE HEAT	1894.00	0.24
TORNADO	1750.00	0.22
HEAT	936.00	0.12
LIGHTNING	722.00	0.09
FLASH FLOOD	658.00	0.08
TSTM WIND	371.00	0.05
RIP CURRENT	271.00	0.03
FLOOD	187.00	0.02
HEAT WAVE	172.00	0.02
RIP CURRENTS	122.00	0.02

And we can also look at the events that caused the most injuries between 1992 and 2011:

```
topinjuries <- data %.% group_by(EVTYPE) %.% summarize(n = sum(INJURIES)) %.%
    mutate(proportion = n/sum(n)) %.% arrange(desc(proportion))
top10injuries <- as.data.frame(head(topinjuries,
    10))
options(xtable.comment = FALSE)
options(xtable.booktabs = TRUE)
xtbl_injuries <- xtable(top10injuries, caption = "Top 10 Events Causing Injuries")
print(xtbl_injuries, floating = FALSE, include.rownames = FALSE)
```

EVTYPE	n	proportion
TORNADO	28702.00	0.52
EXCESSIVE HEAT	6525.00	0.12
TSTM WIND	4683.00	0.08
LIGHTNING	4578.00	0.08
HEAT	2096.00	0.04
FLASH FLOOD	1330.00	0.02
THUNDERSTORM WIND	1107.00	0.02
HAIL	1041.00	0.02
HURRICANE/TYPHOON	933.00	0.02
THUNDERSTORM WINDS	632.00	0.01

According to page 12 of the directive, in an attempt to save space in the data:

"Estimates should be rounded to three significant digits, followed by an alphabetical character signifying the magnitude of the number, i.e., 1.55B for $1,550,000,000. Alphabetical characters used to signify magnitude include "K" for thousands, "M" for millions, and "B" for billions."

However, when we summarize the **PROPDMGEXP** and **CROPDMGEXP** columns, we see that there are identifiers that are not included in the documentation:

```
options(xtable.comment = FALSE)
options(xtable.booktabs = TRUE)
print(xtable(as.data.frame(t(summary(data$PROPDMGEXP))),
    caption = "Property Damage Identifiers"),
    floating = FALSE)
```

	V1	-	?	+	0	1	2	3	4	5	6	7	8	B	h	H	K	m	M
1	303983	0	7	2	141	20	7	3	3	22	3	5	1	25	1	3	259841	1	5597

```
print(xtable(as.data.frame(t(summary(data$CROPDMGEXP))),
    caption = "Crop Damage Identifiers"), floating = FALSE)
```

	V1	?	0	2	B	k	K	m	M
1	396836	2	8	1	7	3	171303	0	1505

Additionally, it looks like some of the identifiers are capitalized and some are not. Because of this, we convert the lower-case identifiers to upper-case. Then we can convert the estimated economic damages to real dollars prior to analyzing the data. To do this we must exponentiate the damages according to their corresponding identifier. We assume that if an identifier is not recognized, the amount of the damage is stated in its nominal terms:

```
data$PROPDMG[which(data$PROPDMGEXP %in% c("h",
    "H"))] <- data$PROPDMG[which(data$PROPDMGEXP %in%
    c("h", "H"))] * 10^2
data$PROPDMG[which(data$PROPDMGEXP %in% c("k",
    "K"))] <- data$PROPDMG[which(data$PROPDMGEXP %in%
    c("k", "K"))] * 10^3
data$PROPDMG[which(data$PROPDMGEXP %in% c("m",
    "M"))] <- data$PROPDMG[which(data$PROPDMGEXP %in%
    c("m", "M"))] * 10^6
data$PROPDMG[which(data$PROPDMGEXP %in% c("b",
    "B"))] <- data$PROPDMG[which(data$PROPDMGEXP %in%
    c("b", "B"))] * 10^9
```

And now we can look at the events that caused the most economic damages between 1992 and 2011:

```
topdamages <- data %.% group_by(EVTYPE) %.% summarize(n = sum(PROPDMG +
    CROPDMG)) %.% mutate(proportion = n/sum(n)) %.%
    arrange(desc(proportion))
top10damages <- as.data.frame(head(topdamages,
    10))
options(xtable.comment = FALSE)
options(xtable.booktabs = TRUE)
xtbl_damages <- xtable(top10damages, caption = "Top 10 Events Causing Economic Damages")
print(xtbl_damages, floating = FALSE, include.rownames = FALSE)
```

EVTYPE	n	proportion
HURRICANE/TYPHOON	58856093646.48	0.29
STORM SURGE	43150368005.00	0.21
TORNADO	23763897052.29	0.12
FLASH FLOOD	13030085264.26	0.06
HURRICANE	10853181769.31	0.05
FLOOD	10163097858.08	0.05
TROPICAL STORM	7597574797.62	0.04
HAIL	6025503023.99	0.03
RIVER FLOOD	5031606060.00	0.02
STORM SURGE/TIDE	4635686000.00	0.02

Now we can plot the top 10 event types for each of fatalities, injuries, and economic damages in a Pareto chart to really *see* which ones are the worst:

```
pareto.chart(xtabs(n ~ EVTYPE, data = top10fatals,
    drop.unused.levels = TRUE), ylab = "Fatalities",
    ylab2 = "Cumulative Percentage", cumperc = seq(0,
        100, by = 20))
```

```
Pareto chart analysis for xtabs(n ~ EVTYPE, data = top10fatals, drop.unused.levels = TRUE)
                 Frequency Cum.Freq.
  EXCESSIVE HEAT      1894      1894
  TORNADO             1750      3644
  HEAT                 936      4580
  LIGHTNING            722      5302
  FLASH FLOOD          658      5960
  TSTM WIND            371      6331
  RIP CURRENT          271      6602
  FLOOD                187      6789
  HEAT WAVE            172      6961
  RIP CURRENTS         122      7083

Pareto chart analysis for xtabs(n ~ EVTYPE, data = top10fatals, drop.unused.levels = TRUE)
```

```
                  Percentage Cum.Percent.
EXCESSIVE HEAT  26.740082     26.74008
TORNADO         24.707045     51.44713
HEAT            13.214740     64.66187
LIGHTNING       10.193421     74.85529
FLASH FLOOD      9.289849     84.14514
TSTM WIND        5.237894     89.38303
RIP CURRENT      3.826062     93.20909
FLOOD            2.640124     95.84922
HEAT WAVE        2.428350     98.27757
RIP CURRENTS     1.722434    100.00000
```

```
abline(h = (sum(top10fatals$n) * 0.8), col = "red",
    lwd = 4)
```

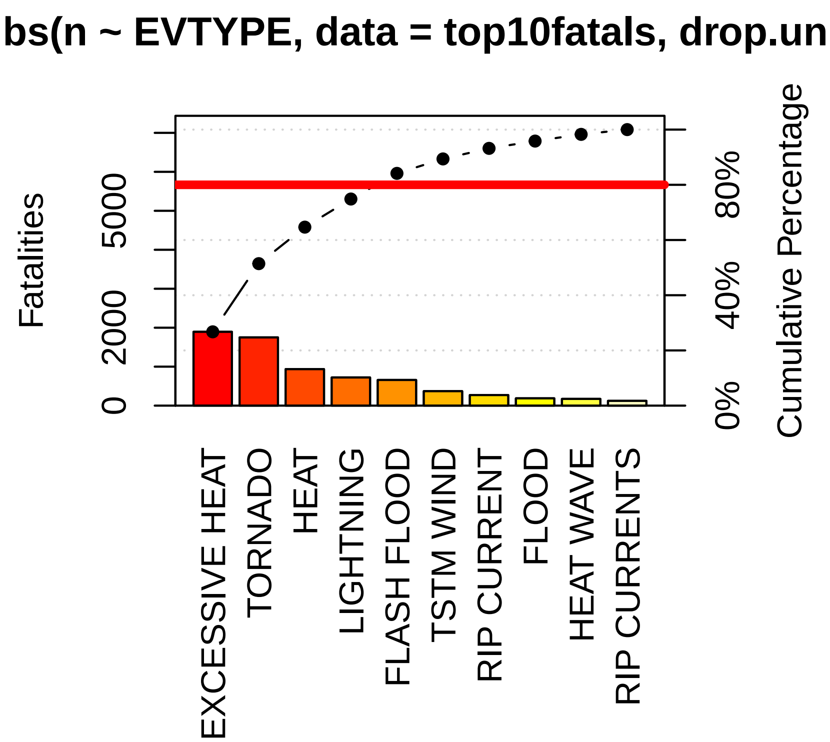


Figure 1: Pareto Chart for Events Causing Fatalities

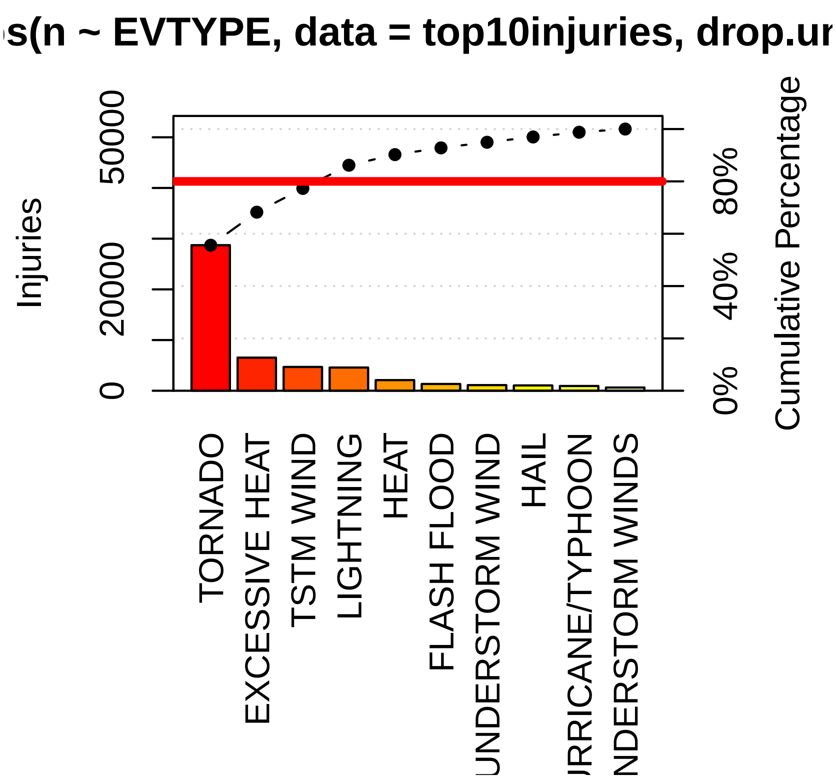
```
pareto.chart(xtabs(n ~ EVTYPE, data = top10injuries,
    drop.unused.levels = TRUE), ylab = "Injuries",
    ylab2 = "Cumulative Percentage", cumperc = seq(0,
        100, by = 20))
```

```
Pareto chart analysis for xtabs(n ~ EVTYPE, data = top10injuries, drop.unused.levels = TRUE)
                     Frequency Cum.Freq.
  TORNADO                28702     28702
  EXCESSIVE HEAT          6525     35227
  TSTM WIND               4683     39910
  LIGHTNING               4578     44488
  HEAT                    2096     46584
  FLASH FLOOD             1330     47914
  THUNDERSTORM WIND       1107     49021
  HAIL                    1041     50062
  HURRICANE/TYPHOON        933     50995
  THUNDERSTORM WINDS       632     51627

Pareto chart analysis for xtabs(n ~ EVTYPE, data = top10injuries, drop.unused.levels = TRUE)
                     Percentage Cum.Percent.
  TORNADO             55.594941     55.59494
  EXCESSIVE HEAT      12.638736     68.23368
  TSTM WIND            9.070835     77.30451
  LIGHTNING            8.867453     86.17196
  HEAT                 4.059891     90.23186
  FLASH FLOOD          2.576171     92.80803
  THUNDERSTORM WIND    2.144227     94.95225
  HAIL                 2.016387     96.96864
  HURRICANE/TYPHOON    1.807194     98.77583
  THUNDERSTORM WINDS   1.224166    100.00000
```

```
abline(h = (sum(top10injuries$n) * 0.8), col = "red",
    lwd = 4)

pareto.chart(xtabs(n ~ EVTYPE, data = top10damages,
    drop.unused.levels = TRUE), ylab = "Damages (in $'s)",
    ylab2 = "Cumulative Percentage", cumperc = seq(0,
        100, by = 20))
```

```
Pareto chart analysis for xtabs(n ~ EVTYPE, data = top10damages, drop.unused.levels = TRUE)
                      Frequency    Cum.Freq.
  HURRICANE/TYPHOON 58856093646  58856093646
  STORM SURGE       43150368005 102006461651
  TORNADO           23763897052 125770358704
  FLASH FLOOD       13030085264 138800443968
  HURRICANE         10853181769 149653625737
  FLOOD             10163097858 159816723595
  TROPICAL STORM     7597574798 167414298393
```

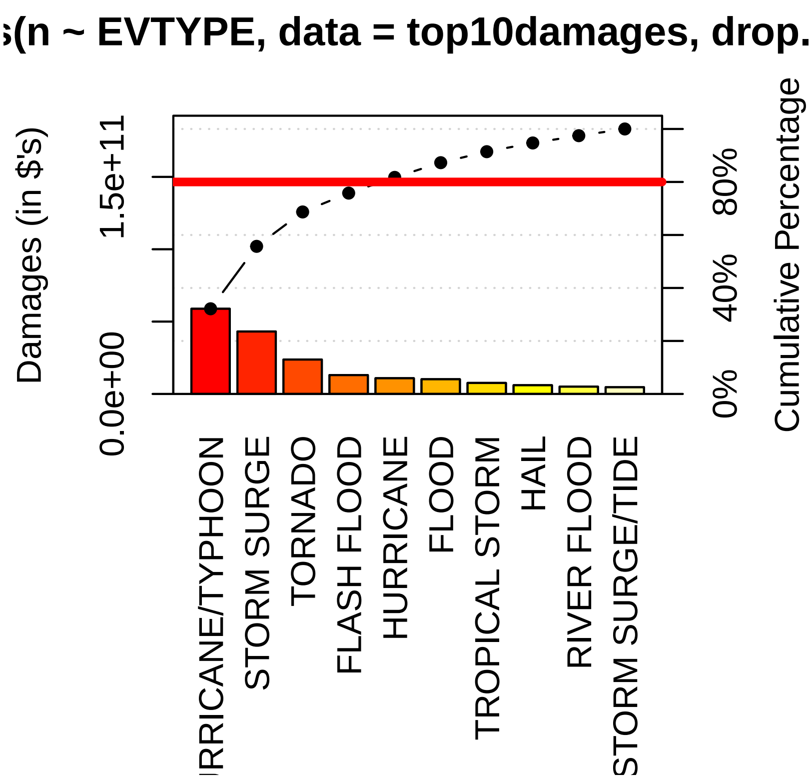
Figure 2: Pareto Chart for Events Causing Injuries



```
  HAIL                6025503024 173439801417
  RIVER FLOOD         5031606060 178471407477
  STORM SURGE/TIDE    4635686000 183107093477

Pareto chart analysis for xtabs(n ~ EVTYPE, data = top10damages, drop.unused.levels = TRUE)
                    Percentage Cum.Percent.
  HURRICANE/TYPHOON  32.142989    32.14299
  STORM SURGE        23.565645    55.70863
  TORNADO            12.978141    68.68678
  FLASH FLOOD         7.116101    75.80288
  HURRICANE           5.927232    81.73011
  FLOOD               5.550357    87.28047
  TROPICAL STORM      4.149252    91.42972
  HAIL                3.290699    94.72042
  RIVER FLOOD         2.747903    97.46832
  STORM SURGE/TIDE    2.531680   100.00000
```

```r
abline(h = (sum(top10damages$n) * 0.8), col = "red",
    lwd = 4)
```

Figure 3: Pareto Chart for Events Causing Economic Damages



Results

Between 1992 and 2011, **EXCESSIVE HEAT** is the event type that caused the most fatalities, **TORNADO** is the event type that caused the most injuries, and **HURRICANE/TYPHOON** is the event type that caused the most economic damages to property and crops.