Title: NOAA Storm Analysis

In our analysis, we were interested in exploring the overall aggregate statistics of NOAA Storm data, specifically the locations (states), beginning time of day, and overall impact through the years of these events.

Our analysis suggests Texas is the overall worst location, based on sheer number of events, and that the overall trend of 'impacts' (defined as the total sum of deaths and injuries) has been increasing over the years.

We also highlight some potentially interesting areas for further analysis, such as; exploring potentially interesting outliers in the reported time of day the event was reported as starting, as well as considering the value of normalizing the 'impact' measure against the overall number of events in a particular state.

Synopsis:

Our analysis includes a number of implicit and explicit data assumptions, particularly around the minimal impact unavailable (`NA`) and inconsistent/invalid data (e.g. non-specific timezone specifications). We also did not normalize for timezone (e.g. such as using `GMT` or daylight savings (`DST`), based on a realistic assumption that all events seem to be U.S.-centric (one possible exception might be in events that cross over international boundaries). However, the largest theoretically skew (between Eastern and Pacific timezones) could be as great as 8 or more (depending on DST), so this may bear further considration.

As we can see from the analysis, the overall impact (defined in our analysis as the sum of deaths and injuries) does appear to be going on a yearly basis. We also learned that being located Texas is a major risk factor. One potential future exploration would be to normalize impact by number of events and determine if states had 'learned' ways to control the overall impact, even in areas where there were high numbers of incidents.

We also can see that there is a definite periodicity in the reported beginning times of events, with a few suspicious outliers, particuarly around what looks to be midday (12:00). This is likely driven more by human behavior than natural ones - particularly given the obvious dip in reported events around that time. Thus implying that either a lot of events are reported as 'noon', when they really start +|- that time, or that is a human-driven reporting artifact

Background

This the R Markdown document associated with my repository located at:

https://github.com/thecapacity/RepData_PeerAssessment2

This document will be used to capture the results of my data analysis in order to make them reproducable, and will be published at my RPubs Account. This document will represent a stand alone assessment, but for more details please check out the GitHub Repository.

Per advice from the instructor, this analysis has been loosely modeled off the example located here: http://www.rpubs.com/rdpeng/13396

My analysis is also published at the following RPubs Location.

The final output for this assignment will be generated via the console with: `knit2html("NOAA_Storm_Analysis.Rmd")`.

```
### Setup some defaults

# Ensure pristine working environment
## rm(list=ls())
## This has been commented out for submission to ensure no disruptivie s

library(knitr)
opts_chunk$set(echo = TRUE, fig.path="figure/", dev="png")
options(scipen = 1) # Turn off scientific notations for numbers

# Load utility libraries
library(data.table)
```

```
## data.table 1.9.4  For help type: ?data.table
## *** NB: by=.EACHI is now explicit. See README to restore previous beh
```

These global defaults are set, or suggested `# As comments` to promote consistent behavior.

This work was done on a Macbook, running OSX 10.9 with the software stack summarized as follows:

```
# Summarize the analysis environment
version
```

```
##                _
## platform       x86_64-apple-darwin13.4.0
## arch           x86_64
## os             darwin13.4.0
## system         x86_64, darwin13.4.0
## status
## major          3
## minor          1.2
## year           2014
## month          10
## day            31
## svn rev        66913
## language       R
## version.string R version 3.1.2 (2014-10-31)
## nickname       Pumpkin Helmet
```

```
sessionInfo()
```

```
## R version 3.1.2 (2014-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] data.table_1.9.4 knitr_1.8
##
## loaded via a namespace (and not attached):
##  [1] chron_2.3-45     digest_0.6.6     evaluate_0.5.5   formatR_1.0
##  [5] htmltools_0.2.6  plyr_1.8.1       Rcpp_0.11.3      reshape2_1.4.
##  [9] rmarkdown_0.3.10 stringr_0.6.2    tools_3.1.2      yaml_2.1.13
```

This fully analysis assumes the `bzunzip2` and `wc` commands are available to extract the data via the command line.

Data Processing

This section outlines (in words and code) how the data were loaded into R and processed

for subsequent analysis.

Analysis will start from the raw CSV file containing the data; and there will be no (pre)processing outside of this document.

As some preprocessing is time-consuming the cache = TRUE option may be used for certain code chunks.

```
## Data Processing Code is here, to load and format the data
##      Subsequent analysis, actually deriving results is captured in th

data_url <- "https://d396qusza40orc.cloudfront.net/repdata%2Fdata%2FStor
data_file <- "data_dir/data"
dateDownloaded <- date()

if (! file.exists("data_dir")) {
    dir.create("data_dir")
    bz_data_file <- "./data_dir/data.bz2"
    download.file(data_url, mode="wb", destfile=bz_data_file, method="cu
    system2("bunzip2", args=c("-dfq", "data_dir/data.bz2") )
}
```

```
# Note, data is extracted every time - but not necessarially downloaded
if ( file.exists(data_file)) {
    my_data <- data.table( read.csv(data_file) )
    # Not strictly necessary, but kept for documentation/completness
    setnames(my_data, make.names( names(my_data) ) )
}

# The first few lines for this data file are:
readLines(data_file, 3)
```

```
## [1] "\"STATE__\",\"BGN_DATE\",\"BGN_TIME\",\"TIME_ZONE\",\"COUNTY\",\
## [2] "1.00,4/18/1950 0:00:00,\"0130\",\"CST\",97.00,\"MOBILE\",\"AL\",
## [3] "1.00,4/18/1950 0:00:00,\"0145\",\"CST\",3.00,\"BALDWIN\",\"AL\",
```

```
# The first few lines of the data read are:
head(my_data, 3)
```

```
##    STATE__           BGN_DATE BGN_TIME TIME_ZONE COUNTY COUNTYNAME STA
## 1:       1 4/18/1950 0:00:00     0130       CST     97     MOBILE
## 2:       1 4/18/1950 0:00:00     0145       CST      3    BALDWIN
## 3:       1 2/20/1951 0:00:00     1600       CST     57    FAYETTE
##     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
##    COUNTYENDN END_RANGE END_AZI END_LOCATI LENGTH WIDTH F MAG FATALIT
## 1:         NA         0                      14.0   100 3   0
## 2:         NA         0                       2.0   150 2   0
## 3:         NA         0                       0.1   123 2   0
##    INJURIES PROPDMG PROPDMGEXP CROPDMG CROPDMGEXP WFO STATEOFFIC ZONE
## 1:       15    25.0          K       0
## 2:        0     2.5          K       0
## 3:        2    25.0          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
```

The data used for this analysis was downloaded on `Sun Jan 25 16:48:05 2015`.

Some information on the overall file length is:

```
# The total line count for this file is:
system2("wc", args=c("-l", data_file), stdout = TRUE)
```

```
## [1] " 1232705 data_dir/data"
```

Note, the above command will likely only work on a Unix-like platform.

R reads `902297` total observations, for an object size of `4.293384 &times; 10<sup>8</sup>`.
The alternate command `read.csv(data_file, comment.char = "#", na.strings = "")` was tried with identical results.
During this analysis, `object.size(my_data)` was `4.293384 &times; 10<sup>8</sup>`.

Note, the following activities were considered - **but not conducted** - to clean/augment the data:

- Instances of “CDT” for `TIME_ZONE` could be changed to `CDT6CST`: This is because R does not recognize the string “CDT” as a valid timezone on my platform.
- Times could be better parsed, e.g. is “15”: But was not done because it's unclear whether that is supposed to be 00:15 (15 past midnight) or 15:00 (3 PM).

However, these activities were not deemed strictly necessary for our analysis and mentioned only for completeness of documenting assumptions.

Data Summarization

This subsection of our processing activities captures the data summization activities conducted for subsequent analysis.

```
summary_data <- data.table( DATE=my_data$BGN_DATE, TIME=my_data$BGN_TIME
summary_data$DATE <- as.character(summary_data$DATE)
summary_data$TIME <- as.factor(summary_data$TIME)

summary_data$IMPACT <- my_data[, FATALITIES] + my_data[, INJURIES]
summary_data$MONTH <- as.numeric(sapply(summary_data[, DATE], FUN=funct
summary_data$YEAR <- as.numeric(sapply(summary_data[, DATE], FUN=functi
```

Analysis

Missing values may cause subtle problems so we check to se what proportion of the observations are missing (i.e. coded as NA).

```
mean( is.na(my_data) )
```

```
## [1] 0.05229737
```

Because the proportion of missing values is low (`0.05229737` in our analysis), **we choose to ignore missing values for now**.

Data variability might also be a problem (e.g. misspellings, etc) however for this analysis we assume minimal expected impact from those potential variations due to the questions being addressed. Also, because the number and type of events have changed over the years our analysis will focus on questions for the totality of data, i.e. regardless of event type.

Specifically we are interested in insights to answer the following three (3) questions:

1. In which state(s) are events most likely to occur?

   It seems interesting to attempt to discern where the most 'dangerous' locations are, and if there is locality consistency across events.

2. At which time(s) of day (`AM`|`PM`) are events most likely to occur?

   *Although events occur across timezones (and possibly daylight savings) the expected +/- caused by this **will be ignored. This is due to our assumption on data quality (e.g. some timezones are set to CDT, which R does not recognize - so we would be forced to assume DST), and the expectation that DST (e.g. +l- 1 hr) will have very little (and symmetric) "crossover" (i.e. leaving an event in AM when it should be in PM or vice versa.*

3. Have the overall `impact` of the events (i.e. `Impacts = Fatalities + Injuries`) increased over time (i.e. each year)?

   Again, focusing on the overall data we attempt to develop a more emperical understanding of the significance of events.

1. First, let us look at the total incidents by state:

```
incidents_by_state <- table(summary_data$STATE)

incidents_by_state
```

```
##
##    AK    AL    AM    AN    AR    AS    AZ    CA    CO    CT    DC
##  4391 22739  1879  3250 27102   257  6156 10780 20473  3294   437  19
##    FL    GA    GM    GU    HI    IA    ID    IL    IN    KS    KY
## 22124 25259  5337   306  2547 31069  4767 28488 21506 53440 22092 173
##    LC    LE    LH    LM    LO    LS    MA    MD    ME    MH    MI
##   274  1526   654  1347    70   262  5652  8185  4524     1 17910 236
##    MO    MS    MT    NC    ND    NE    NH    NJ    NM    NV    NY
## 35648 22192 14695 25351 14632 30271  3022  8075  7129  3139 21058 249
##    OK    OR    PA    PH    PK    PM    PR    PZ    RI    SC    SD
## 46802  4821 22226    28    23     1  3015    96   839 17126 21727
##    ST    TN    TX    UT    VA    VI    VT    WA    WI    WV    WY
##     1 21721 83728  4135 21189   338  3871  3312 19781  9099  7332
```

In our analysis, every state had at least `1` reported incident with the maxium number being `83728`, which belonged to `TX`.
Apparently R prints the 'key' on the console, with `sort(incidents_by_state,`

`decreasing=TRUE)[1]` knitr will not.

2. Next, let us look at the overall start time for events captured.

```
incidents_by_start_time <- table(summary_data$TIME)

# Start Time summary not printed due to the large number of results.
## incidents_by_start_time
```

The greatest number of events for a single time (`CST`) was: `10163`.

The with no time factor having less than `1` event.

Again, R makes it hard to find the 'key' associated with the major|minor value, so they are left as an exercise to the reader.

3. Finally, let us look at the overall impact by year:

```
impact_by_year <- summary_data[, sum(IMPACT), by = YEAR]

impact_by_year
```

```
##     YEAR    V1
##  1: 1950   729
##  2: 1951   558
##  3: 1952  2145
##  4: 1953  5650
##  5: 1954   751
##  6: 1955  1055
##  7: 1956  1438
##  8: 1957  2169
##  9: 1958   602
## 10: 1959   792
## 11: 1960   783
## 12: 1961  1139
## 13: 1962   581
## 14: 1963   569
## 15: 1964  1221
## 16: 1965  5498
## 17: 1966  2128
## 18: 1967  2258
## 19: 1968  2653
## 20: 1969  1377
## 21: 1970  1428
## 22: 1971  2882
## 23: 1972  1003
## 24: 1973  2495
## 25: 1974  7190
## 26: 1975  1517
## 27: 1976  1239
## 28: 1977   814
## 29: 1978   972
## 30: 1979  3098
## 31: 1980  1185
## 32: 1981   822
## 33: 1982  1340
## 34: 1983   853
## 35: 1984  3018
## 36: 1985  1625
## 37: 1986   966
## 38: 1987  1505
## 39: 1988  1085
## 40: 1989  1754
## 41: 1990  1920
## 42: 1991  1428
```

```
## 43: 1992  1808
## 44: 1995  5971
## 45: 1994  4505
## 46: 1993  2447
## 47: 1996  3259
## 48: 1997  4401
## 49: 1998 11864
## 50: 1999  6056
## 51: 2000  3280
## 52: 2001  3190
## 53: 2002  3653
## 54: 2003  3374
## 55: 2004  2796
## 56: 2005  2303
## 57: 2006  3967
## 58: 2007  2612
## 59: 2008  3191
## 60: 2009  1687
## 61: 2010  2280
## 62: 2011  8794
##     YEAR    V1
```

The year with the greatest number of injuries and deaths is: `1998`.

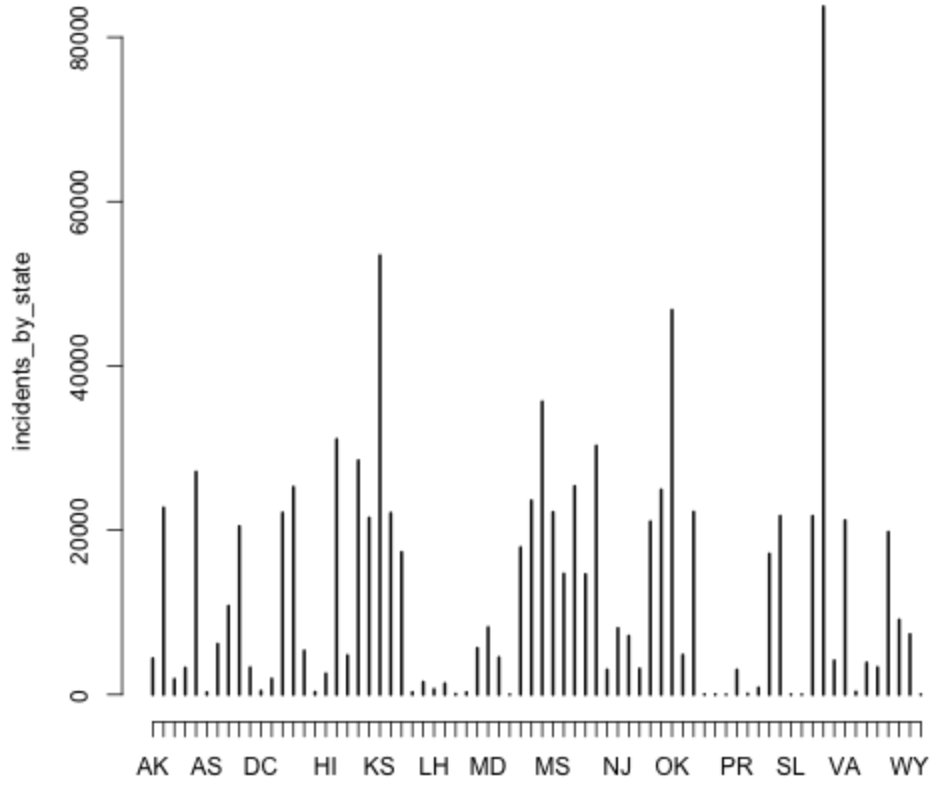
The year with the least number of injuries and deaths is: `1951`.

Results

This section will present the final results. Only final graphs and a discussion of conclusions will be captured here, with all computational work being done in the earlier sections above.
This section has at least one figure containing a plot, but no more than three figures.
Per assignment guidance, figures may have multiple plots in them (i.e. panel plots), but there will not be more than three figures total.

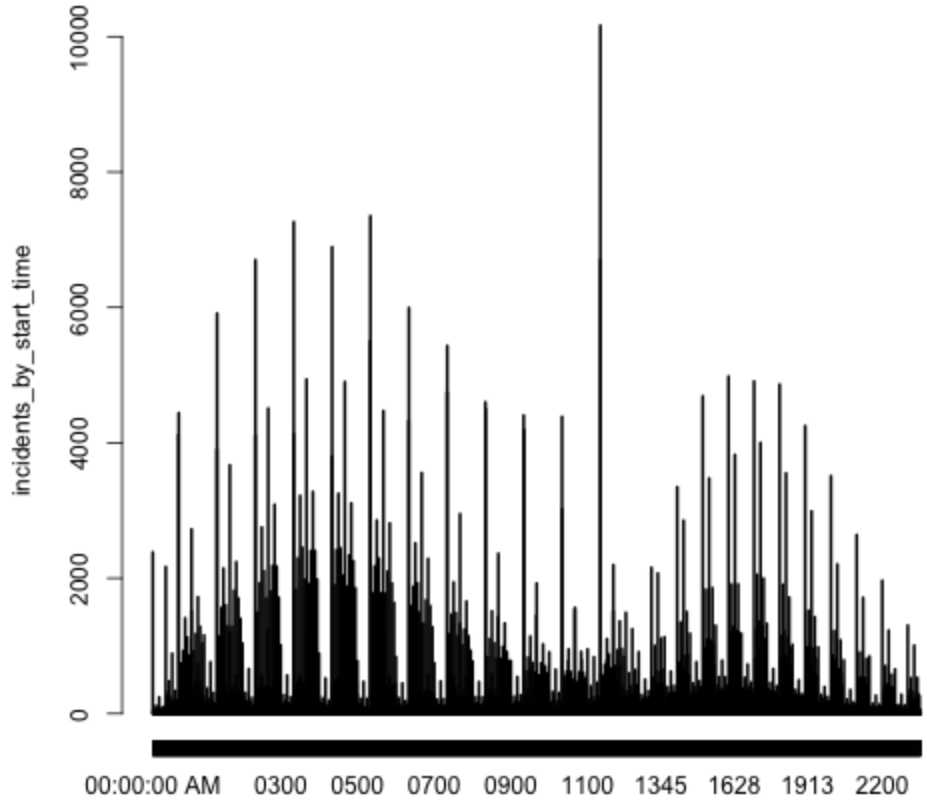
1. Incidents by State

```
plot(incidents_by_state)
```



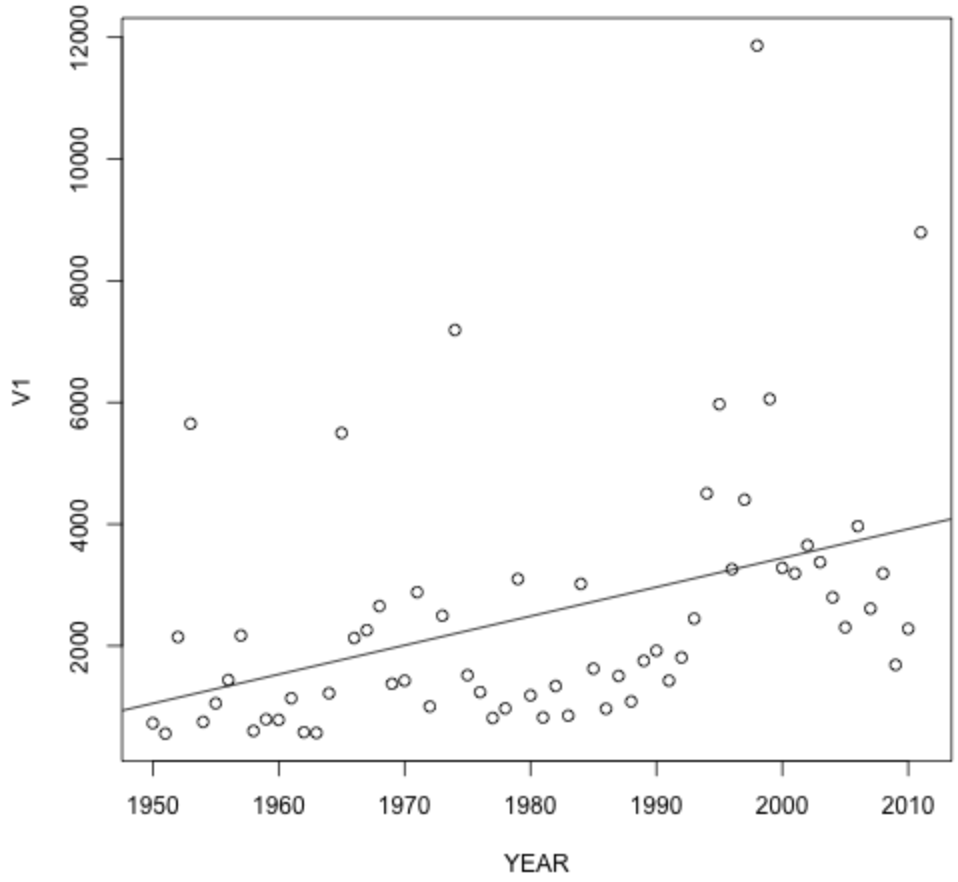
2. Occurances by Start Time

```
plot(incidents_by_start_time)
```



3. Yearly Impact

```
plot(impact_by_year)
abline(lm(impact_by_year$V1 ~ impact_by_year$YEAR))
```



As we can see from the analysis, the overall impact (defined in our analysis as the sum of deaths and injuries) does appear to be going on a yearly basis. We also learned that being located Texas is a major risk factor. One potential future exploration would be to normalize impact by number of events and determine if states had 'learned' ways to control the overall impact, even in areas where there were high numbers of incidents.

We also can see that there is a definite periodicity in the reported beginning times of events, with a few suspicious outliers, particuarly around what looks to be midday (12:00). This is likely driven more by human behavior than natural ones - particularly given the obvious dip in reported events around that time. Thus implying that either a lot of events are reported as 'noon', when they really start +l- that time, or that is a human-driven reporting artifact.