AirSafe.com traffic spikes heat maps May 2006 to November 2015

Todd Curtis

November 22, 2015

Summary

A previous AirSafe.com study, AirSafe.com traffic spikes May 2006 to November 2015, reviewed traffic to the web site and determined that a total of 266 days over that over nine and one half year time span where the estimated number of visits to the site was at least two standard deviations higher than the average number of visits during a 28-day comparison period. The previous study showed that most of these events, called traffic spikes, were associated with identifiable aerospace-related events, and that spike days were more likely during some days of the week. This study used the same data from the previous study to create visual summaries of the traffic spike data that would help to further illustrate traffic spike patterns associated with the day of the week and to show if there were also clear patterns showing up in longer time scales.

Introduction

The web site AirSafe.com, which has been in operation since July 1996, provides the aviation safety community and the general public with useful information about aviation safety and security. The site highlights a particular class of events, which it defines as significant events, a category that typically includes events involving airliner deaths or other aerospace events that attract significant amounts news media attention. These significant events all all listed in one or more places on AirSafe.com, but at a minium are listed on summary pages that list the significant events from a given year.

Between 10 May 2006 and 14 November 2015, there were 266 days with significant traffic spikes on AirSafe.com, specifically days where the estimated number of visits to the site exceeded the average number of visit in a 28-day period that begins five weeks before the day being measured, and ends one week prior to the day being measured by two or more standard deviations. For example, the number of visits on 31 October 2015 would be compared with the distribution of the number of visits from the period 27 September 2015 to 24 October 2015.

Traffic on the site is measured by Google Analytics, which uses tracking codes that provided detailed information about how users interact with the web site. The code doesn't track individuals, but rather interactions that Google Analytics defines as a session, which is an interaction between AirSafe.com and some identifiable location or device that is connected to the Internet.

Between May 2006 and November 2015, significant traffic spikes on AirSafe.com, specifically days where the estimated number of visits to the site exceeded the average number of visit in a 28-day comparison period by more than two standard deviations, frequently occurred within seven days of widely reported events involving airline safety and security, but there were a number of spikes that occurred that were unrelated to a recent safety or security event.

Purpose of the analysis

The intent of this analysis is to take the information about spikes, specifically the days that the spike occurs, and to put them in to a visual display that may reveal patterns of spike creation, specifically patterns related to th days of the week or the months of of the year that spikes occur.

Measuring site traffic

Traffic on the site is measured using Google Analytics, which provides detailed information about how users interact with the web site. The code doesn't track the behavior of individuals, but rather interactions that Google Analytics defines as a session, which is an interaction between a web site and some identifiable location or device that is connected to the Internet.

On AirSafe.com, sessions are identified by one or more interactions with the web site, where either a single interaction is followed by 30 minutes of inactivity, or where multiple interactions, for example visits to different pages, that are separated by fewer than 30 minutes.

There is not necessarily a one to one relationship between a session and the actions of the entity or individual responsible for the interaction that results in a session. The tracking code logs a visit from an identifiable location, such as a mobile phone or server. Without further information, it is difficult to determine what or who could be responsible for a single session. For example, the following could all represent one session:

- One person visiting a single page and spending fewer than 10 seconds on the site,
- One person visiting numerous locations on the site over a period of several hours, with less than a 30 minute period of inactivity,
- Several people looking at the same display screen while visiting the site,
- An online application that is automatically visiting web sites.

There is also insufficient information to identify situations where two or more concurrent sessions may actually represent the same person or entity accessing the site on multiple devices, for example, visiting first on a mobile phone before switching to a desktop.

In spite of these and other limitations, two reasonable assumptions were made about session data:

- That a large majority of the session represent the actions of individual people,
- That changes in the number of sessions are highly correlated with changes in the number of individuals accessing the site.

Data

The key source of data were the AirSafe.com session data provided by Google Analytics. The period covered by this study was from 6 April 2006 to 14 November 2015.

```
# Import data (data files online in directory http://www.airsafe.com/analyze/)

sessions.raw = NULL
sessions = NULL
range = 28
# Offset if we want to move the end of the 21 day range to (offset + 1)
# day prior to the day being measured
offset = 7
# Download raw session data
sessions.raw <- read.csv("http://airsafe.com/analyze/sessions.csv", header = TRUE)

# Ensure that working data is in a data frame
sessions = as.data.frame(sessions.raw)
```

The following pre-processing steps were completed prior to the analysis:

- Changing the date format from m/d/yyyy to yyyy-mm-dd
- Changing the column names to “Date” and “Sessions”
- Adding the following columns:
  - date_index - Contains an integer index of days measured, with day one being the first measured day
  - mean_range - Contains the mean value of sessions during the 28-day comparison period
  - sd_range - Contains the standard deviation of the session values from the 28-day comparison period
  - SpikeSD - Standard deviaion of the session values from the measured day compared to the distribution of session values from the comparison period.
  - Spike2mean - Ratio of a particular day’s session value and the mean number of sessions from the comparison period value

The values of these columns were all initialized to the value -1. Based on the session data, the values in each of these five new coloumns, starting with the 29td row, would be updated to reflect the values computed from 28-day comparison period.

```
# Change the column names
colnames(sessions) = c("Date","Sessions")
colnames(sessions.raw) = c("Date","Sessions")

# Convert column of session values from factor to numeric
sessions$Sessions = as.numeric(as.character(sessions.raw$Sessions))
# Dates are in form 5/1/2006, must convert to a date format of yyyy-mm-dd
sessions$Date = as.Date(sessions.raw[,1], "%m/%d/%Y")


# Add columns for the mean and standard deviation of previous defined range of days of
# sessions and give them a default value to aid in identifying days without a spike measurment

sessions$date_index = -1
sessions$mean_range = -1
sessions$sd_range = -1
sessions$SpikeSD = -1
sessions$Spike2mean = -1


# This loop will compute each day's mean, and standard deviation for the
# previous range of days, starting  with the 35th day of data
for(i in (range + offset): nrow(sessions))
{
        sessions$date_index[i] = i-(range + offset) + 1
        sessions$mean_range[i] = mean(sessions$Sessions[(i-(range + offset) + 1):(i-offset)])
        sessions$sd_range[i] = sd(sessions$Sessions[(i-(range + offset) + 1 ):(i-offset)])
        sessions$Spike2mean[i] = sessions$Sessions[i]/sessions$mean_range[i]
        sessions$SpikeSD[i] = (sessions$Sessions[i] - sessions$mean_range[i] )/sessions$sd_range[i]
}
```

Data overview of session values

Because the number of sessions on a particular day had to be compared to the 28-day period that begins 35 days before the day being measured, the first date that could be checked for spikes was 10 May 2006, which was the 35th day after th first date with session data, 6 April 2015.

The initial exploratory data analysis of the session data showed a wide range of values from under 200 to over 75,000 sessions in a particualr day. The following histograms show that that the number of sessions showed a distict postive (rightward skew), however, the log of the session values reveals a much more symmetric distribution.

Because this study was looking at comparisons of a particular day's session values with the distribution of the sample mean of sessions from a subset of the entire range of session values, it was not necessary to model the distribution of the entire population of sessions. It was sufficient to employ the centeral limit theorem and assume a normal distribution of the sample mean of a 28-day sequence of session values.

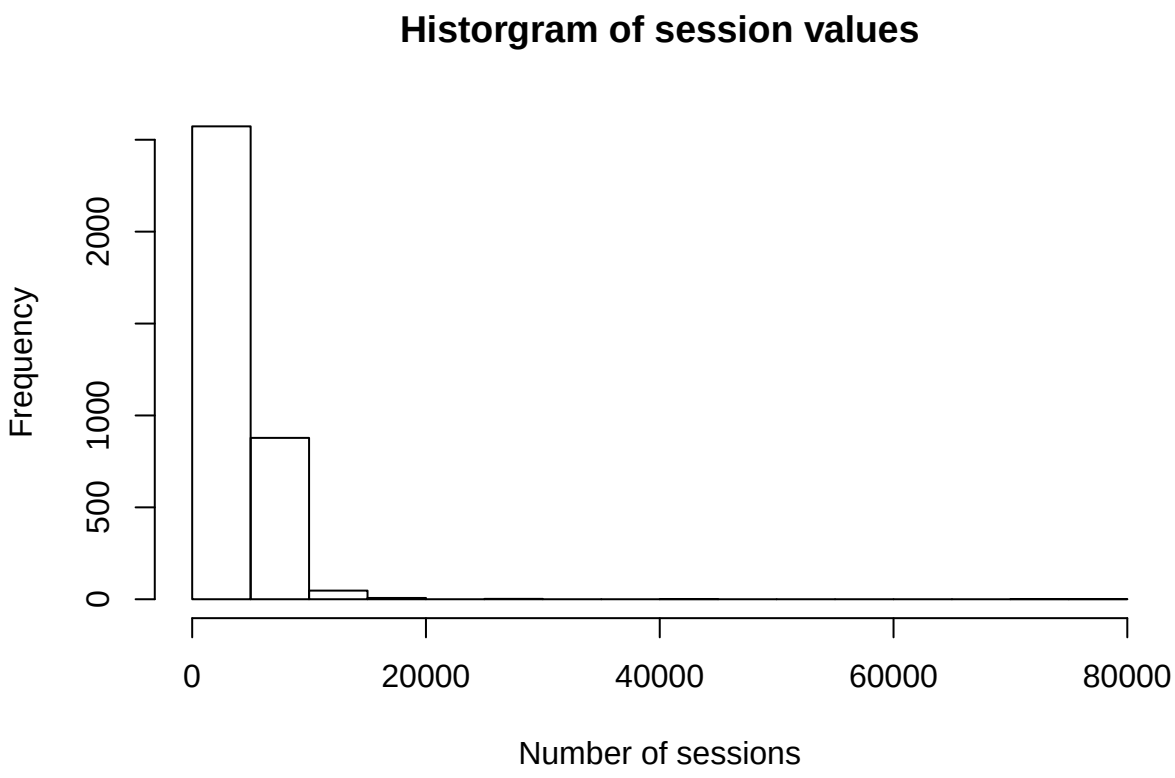
```
# Summary and histogram of the sessions data

summary(sessions$Sessions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     197    2860    3796    4276    5102   75120
```

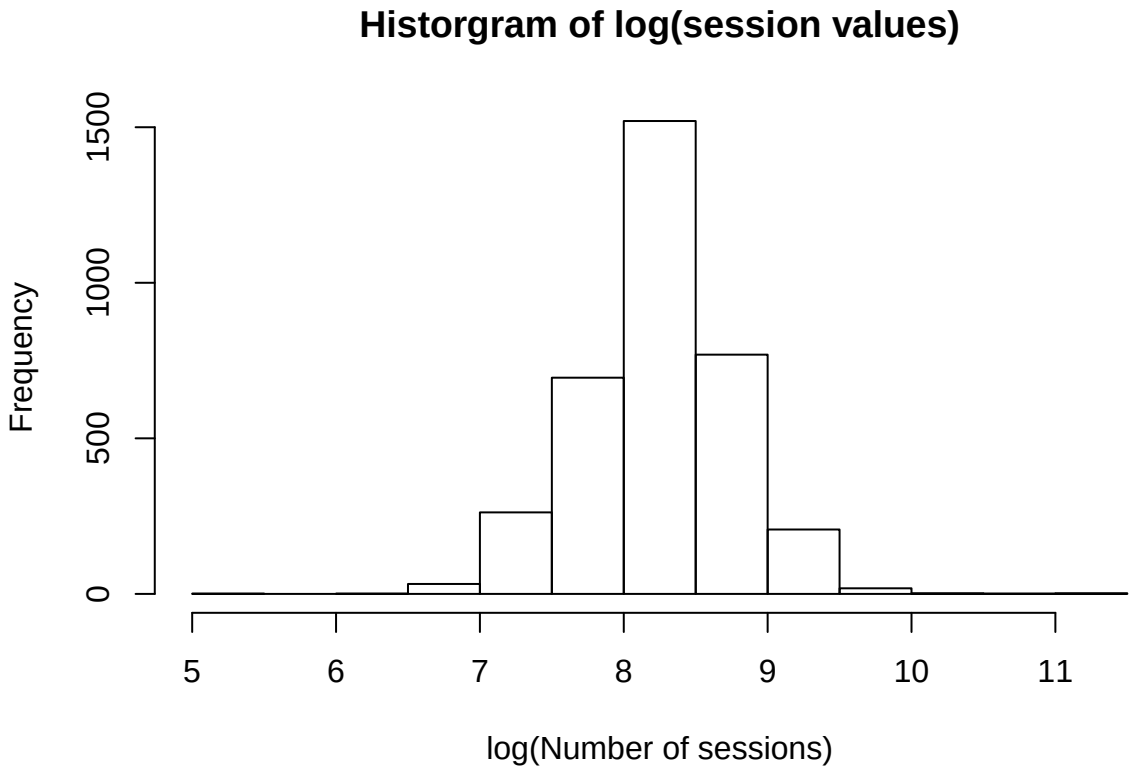
```
summary(log(sessions$Sessions))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.283   7.959   8.242   8.232   8.537  11.230
```

```
hist(sessions$Sessions, main="Historgram of session values",
     xlab="Number of sessions")
```



```
hist(log(sessions$Sessions), main="Historgram of log(session values)",
     xlab="log(Number of sessions)")
```



Identifying spike days

The next steps included adding a logical vector (SpikeSD) for the spike days, defined as those days were the number of sessions was at least two standard deviations higher than the average number of sessions from the comparison period. Also, the days that were measured were tranfered into a separate data frame (measured.days).

The data from the subset of days with a significant spike in sessions was put into a separate data frame. A total of 266 days met this criteria. In addition, three new variables were added representing the day of the week, the month, and the year that the spike occurred (Day, Month, and Year respectively). A redundant variable (Spike) eliminated since all the values in spike.days would be TRUE.

This data frame representing the days that were measured (measured.days) and the days with spikes (spike.days) were archived in CSV files that are available at

- http://www.airsafe.com/analyze/measured_days.csv
- http://www.airsafe.com/analyze/spike_days.csv

What causes spikes The previous study showed that of the 182 of the 266 spike days, or 68.4% of those spike days were associated with an identifiable envent, most of which were aircraft crashes or other events that no known relation to the day of the week, the season of the year, or any other temporal pattern.

It was assumed that for those cases where the spike was associated with a significant event, traffic to the site was driven by one of three things:

- AirSafe.com marketing efforts to drive traffic to particular web site content after a significant aviation safety or security event.
- Web site searches related to a significant event. The site consistently gets about two-thirds of all traffic as the result of searches on major search engines like Google, Bing, and Yahoo, and an increase in the volume of searches for aviation-related online content often follows in the wake of high media interest in the subject in the aftermath of an event.
- From links on other web sites

The remaining 31.6% of the spike days also did not show any clear or consistent pattern of occurrences, and it was unclear what factors mahy hve led to traffic spikes on those days.

Spike days driven by multiple significant events The previous study pointed out that on three of the 266 spike days, there were two signifiant events which generated enough site traffic to independently lead to a spike day. Because this study was looking at the temporal pattern of spike days rather than the pattern of the magnitude of the traffic that led to the spike days, those three days were not treated treated any differently.

```
# Identifies as spike any SpikeSD (rounded to two significant digits) of 2 or more
sessions$Spike =  round(sessions$SpikeSD, digits=2) >= 2

# Transfer only measurable spike days to a new data frame
measured.days = sessions[sessions$mean_range != -1,]

# Transfer only spike days to a new data frame
spike.days = sessions[sessions$Spike==TRUE,]
spike.days$Year = format(spike.days$Date, "%Y")
spike.days$Month = months(spike.days$Date)
spike.days$Day = weekdays(spike.days$Date)

# Redundant "Spike" column eliminated since all the values in spike.days would be TRUE
spike.days$Spike = NULL

write.csv(spike.days, file = "spike_days.csv")
write.csv(measured.days, file="measured_days.csv")
```

Data analysis

Prior to developing heat maps of the spike day, it was necessary to examine numerical summaries of the data to get an idea of what kinds of trends could be identified using a visual depiction like a heat map.

Spikes were determined by the number of standard deviations, in part to deal with the different levels of traffic in different years. While average traffic increased over a time span measured in years, it did not signficantly change over the 28-day time span used as the comparison period used in this study. A summary of the standard deviation values on spike days showed that half of the spike days had a number of sessions was less than 3.43 standard deviaions above the mean number of sessions in their comparison periods, and ranged from a minium of 2.00 to a maximum of 74.98 standard deviations.

Distribution of spike days by day of the week Perhaps the most striking result was the days on when a spike occurred. Friday and Saturday had considerably fewer spike days than the rest of the week. A Chi-square test on the distibution of the spike days by day of the week had a p-value much smaller than 0.05, so one could reject the null hypothesis that spike days were uniformly distributed among the days of the

week. Using the same approach to look at the distribution of spike days by month, a similar null hypotheses for the distributions by month also hada p-value much smaller than 0.05, and would also be rejected.

```
# Summary of distibutions of spike day sessions
summary(spike.days$Sessions)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1638    4304    5808    7429    9230   75120
```

```
# Summary of distibutions of number of standard deviaions on a spike day
round(summary(spike.days$SpikeSD), digits=2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    2.47    3.43    5.71    6.43   74.98
```

```
# Sorting number of spike days by day of the week
sort(table(spike.days$Day), decreasing = TRUE)
```

```
##
##    Monday Wednesday   Tuesday  Thursday    Sunday    Friday  Saturday
##        55        47        46        41        37        25        15
```

```
# Chi-square test for the null hypothesis of a uniform distribution of
# spike days among the days of the week
chisq.test(table(spike.days$Day))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(spike.days$Day)
## X-squared = 30.053, df = 6, p-value = 3.841e-05
```

```
# Chi-square test for the null hypothesis of a uniform distribution of
# spike days among the months of the year
chisq.test(table(spike.days$Month))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  table(spike.days$Month)
## X-squared = 54.481, df = 11, p-value = 9.642e-08
```

Distribution of spike days by month and year Two ways of looking at the distribution of the spike days by month and year is by using one-dimensional summarary tables (one for spikes by month and another for spikes by year), a two-dimensional summary table. Note that the one-dimensional summary tables below orders the results by decreaseing size.

```
# Distribution of spike days by month


# Sorting number of spike days by month of the year
sort(table(spike.days$Month), decreasing = TRUE)
```

```
##
##   January      July      June       May     March  November    August
##        39        38        31        29        27        21        18
##   October  December  February     April September
##        16        15        13        12         7
```

```
# Sorting number of spike days by year
sort(table(spike.days$Year), decreasing = TRUE)
```

```
##
## 2014 2015 2008 2009 2006 2011 2007 2013 2010 2012
##   62   35   26   26   24   24   20   20   17   12
```

```
# For the two dimentsional table, first, order the year
spike.ordered = spike.days[order(spike.days$Year),]

# First make months factors and order them like the calendar
spike.ordered$Month = factor(spike.ordered$Month,levels=c("January",
        "February","March", "April","May","June","July","August","September",
        "October", "November","December"), ordered=TRUE)

# The table as numbers
table(spike.ordered$Month, spike.ordered$Year)
```

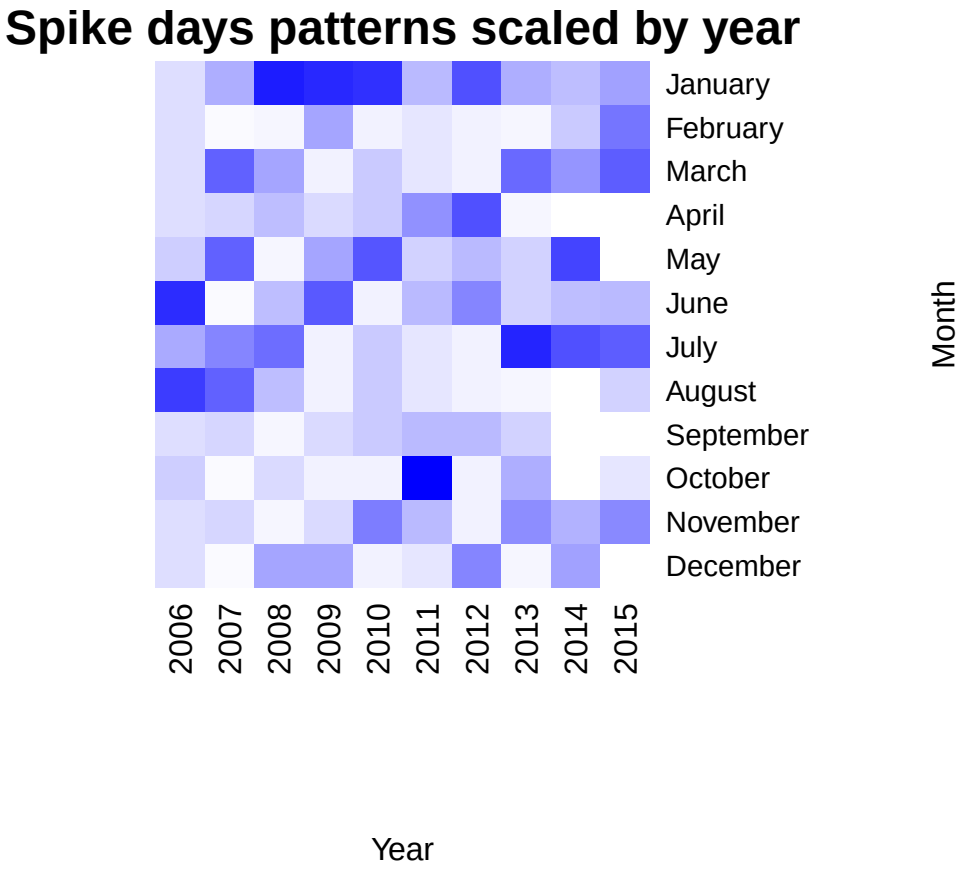
```
##
##             2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   January      0    2    8    8    5    2    3    2    5    4
##   February     0    0    0    3    0    0    0    0    4    6
##   March        0    4    3    0    1    0    0    4    8    7
##   April        0    1    2    1    1    4    3    0    0    0
##   May          1    4    0    3    4    1    1    1   14    0
##   June        10    0    2    6    0    2    2    1    5    3
##   July         3    3    5    0    1    0    0    6   13    7
##   August       9    4    2    0    1    0    0    0    0    2
##   September    0    1    0    1    1    2    1    1    0    0
##   October      1    0    1    0    0   11    0    2    0    1
##   November     0    1    0    1    3    2    0    3    6    5
##   December     0    0    3    3    0    0    2    0    7    0
```

Using heat maps to illustrate the istribution of spike days A third way to display the same information is visually using heat maps. The heat map would use the same data as that in the two-dimensional table, but because of the choice of scaling, serves to emphasize different aspects of the data.

These heat maps use a color scheme where the the lowest values are white and the highest values are dark blue. The first heat map is scaled by the column or x-axis values (Year), which will highlight the months with a realtively high or low number of spikes compared to other months.

Distribution of spike days to emphasize patterns of particular months There were four heat maps created from the underlying data. The first scaled the values for a given year (column values) by the month values across all years in order to show, for a particular month, what years had relatively high or low traffic.

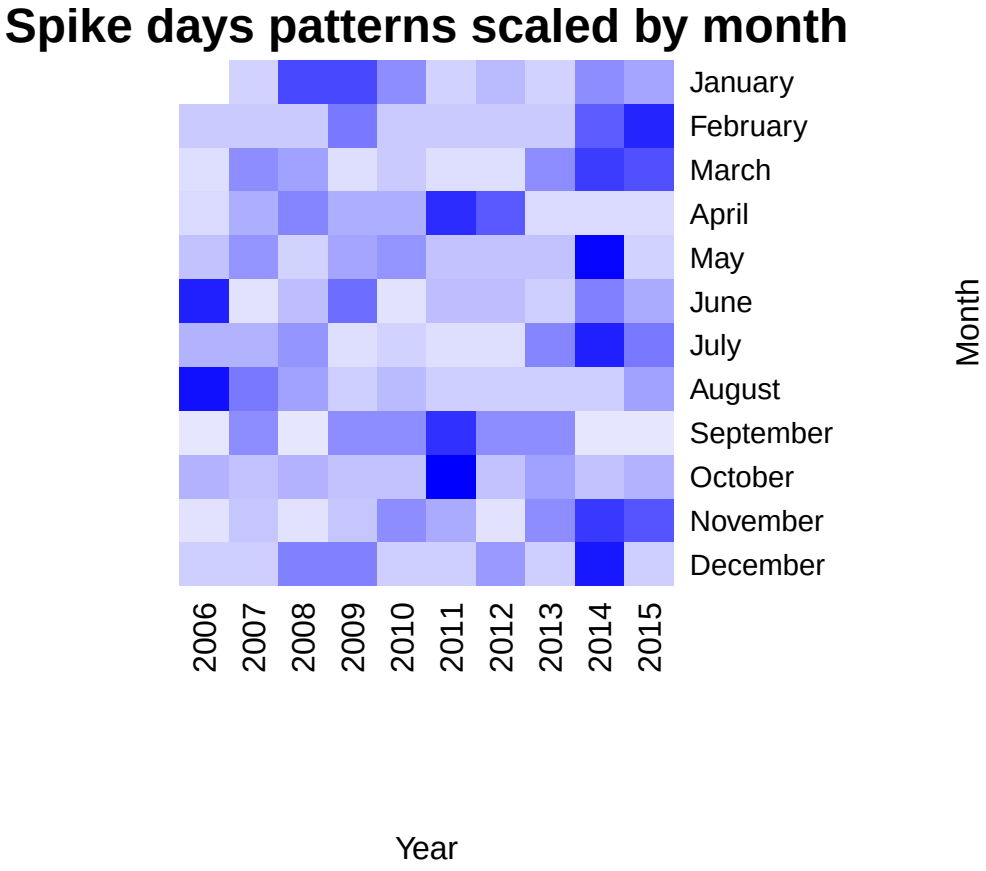
```
# Creating heat map with a cyan to purple color scheme  (option is heat colors using heat.colors)

# When scaling by column (the year), the focus is on the months with a realtively high or low
#  number of spikes compared to other months

# First, create a color pallette that goes from white for lowest to dark blue for the highest value
palette = colorRampPalette(c('#ffffff','#0000ff'))(64)

heatmap(table(spike.ordered$Month, spike.ordered$Year),Rowv=NA, Colv=NA,revC=TRUE,
        scale="column", col = palette, margins=c(9,10),
        main="Spike days patterns scaled by year",
        xlab="Year", ylab="Month")
```



The months of September, November, and December had relatively low numbers of spike days in most years, while January had more years of relatively high numbers of spike days compared to other months.

Distribution of spike days to emphasize patterns of particular year The second heat map scaled the values for a given month (row values) across all years in order to show, for a particular year, what months had relatively high or low traffic.

```
# When scaling by row (the month), the focus is on the years with a realtively high or low
#  number of spikes compared to other months
heatmap(table(spike.ordered$Month, spike.ordered$Year),Rowv=NA, Colv=NA,revC=TRUE,
        scale="row", col = palette, margins=c(9,10),
        main="Spike days patterns scaled by month", xlab="Year", ylab="Month")
```



Distribution of spike days by day of the week Using the same approach as was used to look at the spike day distribution by month, one can also look at the distribution by day of the week. Earlier, it was shown that the distribution of spike days by day of the week was not consistent with a uniform distribution.

The two-dimensional table of spike days by day of the week is below, and shows both a significant number of spike days in 2014 for every day of the week, as well as a low number of spike days on Saturday.

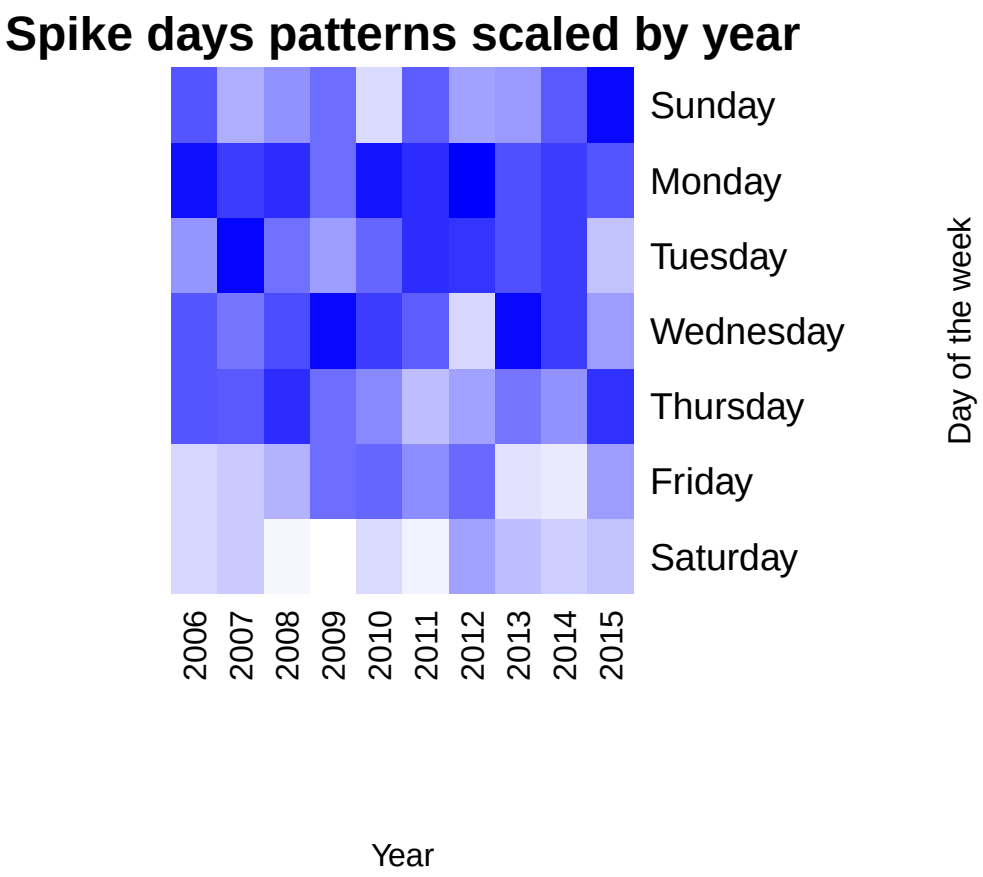
```
# Two-dimensional table of spike days by day of the weej and year

# First make days factors and ordering them with Sunday being the first day
spike.ordered$Day = factor(spike.ordered$Day,levels=c("Sunday","Monday","Tuesday",
        "Wednesday","Thursday","Friday","Saturday"), ordered=TRUE)

# Sorting number of spike days by month of the year
table(spike.ordered$Day,spike.ordered$Year)
```

```
##
##              2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
##   Sunday        4    1    3    4    0    4    1    2   10    8
##   Monday        5    5    6    4    5    5    4    4   11    6
##   Tuesday       3    7    4    3    3    5    3    4   11    3
##   Wednesday     4    3    5    6    4    4    0    6   11    4
##   Thursday      4    4    6    4    2    2    1    3    8    7
##   Friday        2    0    2    4    3    3    2    0    5    4
##   Saturday      2    0    0    1    0    1    1    1    6    3
```

Distribution of spike days to emphasize patterns of particular days The third heat map is scaled by the column or x-axis values (Year), which will highlight, for a particular day of the week, what years had relatively high or low traffic compared to other days of the week.

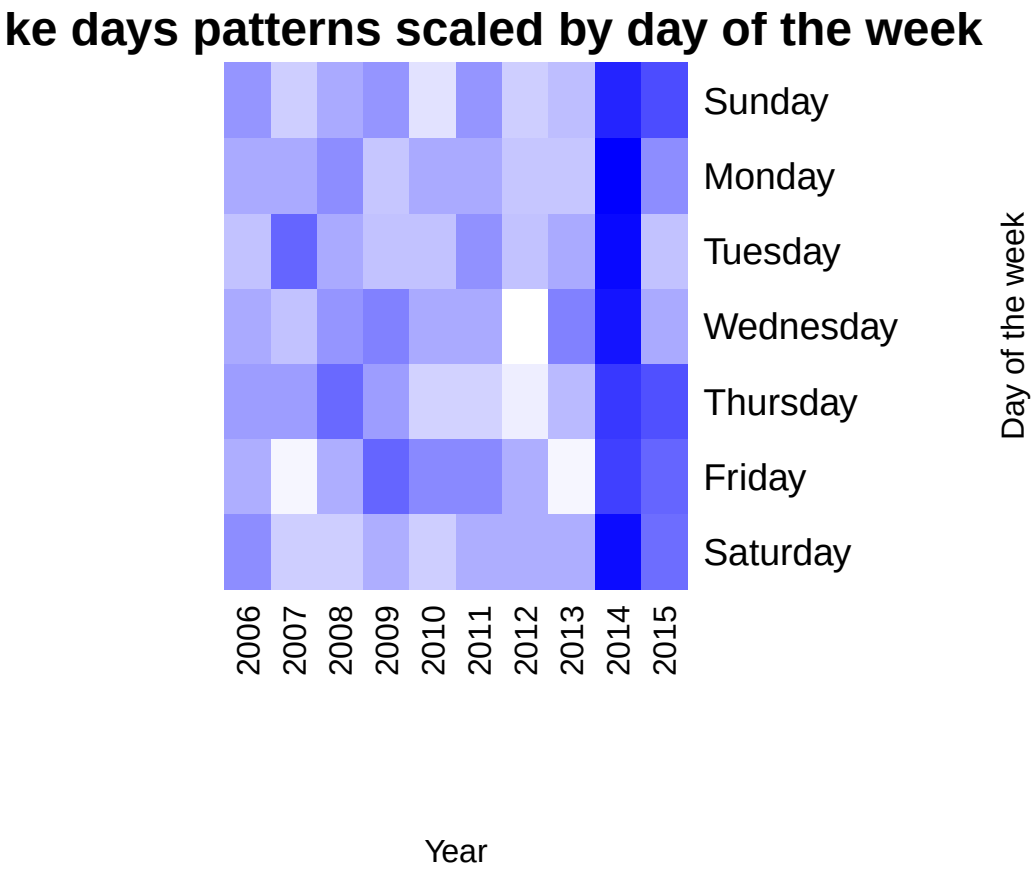
This heat map clearly shows that Monday, Tuesday, and Wednesday have a relatively high proportion of spike days, and Friday and Saturday much less.

```
# Create a heat map for days of the week by ordering them with Sunday being the first day
spike.ordered$Day = factor(spike.ordered$Day,levels=c("Sunday","Monday",
        "Tuesday", "Wednesday","Thursday","Friday","Saturday"), ordered=TRUE)

# Scaling by column (the year) will highlight the days with a relatively
# high or low number of spikes

heatmap(table(spike.ordered$Day, spike.ordered$Year),Rowv=NA, Colv=NA,revC=TRUE,
        scale="column", col = palette, margins=c(9,11),
        main="Spike days patterns scaled by year",
        xlab="Year", ylab="Day of the week")
```



Distribution of spike days to emphasize day of the week patterns of particular years Scaling by row (day of the week) will emphasize the years that have a high number of spike days for a particular day of the week, and 2014 clearly stands out for a relatively high number of spikes for every day of the week, and 2012 has a relatively low number for all the days of the week.

```r
#  Heatmap of spikes by scaled by row (the day of the week) would highlight years with a
#  realtively high or low number of spikes compared to other years. Clearly 2014 is the winner.

heatmap(table(spike.ordered$Day,spike.ordered$Year),Rowv=NA, Colv=NA,revC=TRUE,
        scale="row", col = palette, margins=c(9,11),
        main="Spike days patterns scaled by day of the week",
        xlab="Year", ylab="Day of the week")
```



Findings

While the four generated heat maps used the same underlying raw data, depending on how the heat map was scaled, it would show different temorpar patterns with respect to which days of the week or months of the year were either more or less likely to have a spike day.

The patterns found in the earlier paper with repsect to the days of the week that were more likely to have spike days was clearly shown in the third heat map, with Monday, Tuedsay, and Wednesday consitently winning out over the other days of the wee.

A less striking pattern was evident in the first heat map which showed that September to December had relatively fewer spike days compared to the rest of the year, but no month showed a pattern of relatively high numbers of spikes that was consitent over all of the years of the study.

The second and fourth heat maps, which were both scaled to see if particular years had relativey high or low numbers of spikes, showed that 2014, and to a lesser extent 2015, had generall higher levels of spike days compared to other days of the week.

Data and output The study, as well as the raw and processed data used by the study, are available online:

- Raw data - http://www.airsafe.com/analyze/sessions.csv
- Spike days - http://www.airsafe.com/analyze/spike_days.csv
- Full analysis (PDF) - http://www.airsafe.com/analyze/traffic_heatmaps.pdf
- Full analysis (Rmd) - http://www.airsafe.com/analyze/traffic_spikes.Rmd
- Full analysis (HTML) - http://www.airsafe.com/analyze/traffic_spikes.html