AirSafe.com traffic spikes May 2006 to November 2015

Todd Curtis

November 16, 2015

Summary

Between 10 May 2006 and 14 November 2015, there were a total of 269 significant traffic spikes on AirSafe.com, specifically days where the estimated number of visits was two standard deviations higher than the average number of visits during a comparison period. Most of the spike events were associated with identifiable events:

- 56.1% of the spike events were associated with AirSafe.com-desigated significant aviation events that occurred during the study period.
- 62.5% of the spike events were associted with AirSafe.com-desigated significant aviation events that occurred either before or during the study period.
- 68.8% of the spike events were associted some identifiable aerospace-related event.

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.

While most of these spikes occurred within seven days of widely reported aerospace-related events, some spikes can occur months or even years after an event, and 31.2% of the spikes (84 of 269) can't be associated with a particular event.

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.

While average annual traffic on the site has increased by several orders of magnitude over the past 19 years, two traffic trends have been consistent. The first trend is that normal traffic follows roughly the typical North American work schedule, with traffic during non-holiday weekdays being somewhat higher than weekend traffic, but with a relatively small (less than 50%) differences between the traffic on weekends and on weekends. On some days, traffic on the site experiences significant increases, which are often, but not always, associated with airline or aerospace related events that attract the attention of major news media organizations.

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 look at site traffic for a particular time period to examine the relationship between traffic surges and AirSafe.com-defined significant events in order determine several things:

- Which significant events were associated with a large increase in traffic.
- Which dates had significant increases in traffic that were apparently unrelated with the dates of significant events.

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.

Traffic distribution The distribution of traffic reflects both the dominant role of the US with respect to generating Internet traffic and the fact that the site is in English. Over the study period of over 9.5 years, representing just over 15 million sessions, the top five countries generating sessions on AirSafe.com were:

1. United States: 55.2%
2. United Kingdom: 9.4%
3. Canada: 5.4%
4. Australia: 3.7%
5. India: 2.1%

Data

The two key sources of data were the AirSafe.com session data provided by Google Analytics, and the listings of significant airline safety and security events listed on AirSafe.com. The period covered by this study was from 6 April 2006 to 14 November 2015.

While sessions are defined elsewhere in this report, a significant event is defined in detail at http://www.airsafe.com/events/define.htm, but they typically involve either the deaths of one or more persons during an airline flight, or circumstances involving some aspect of aerospace that led to extensive coverage by major news organizations.

While AirSafe.com has pages summarizing the significant events in a particular year starting from 1996, the significant events from the period covered by this study were of particular interest. The significant event data for the years 2006-20015 on AirSafe.com were located on the following pages:

- 2006 - http://www.airsafe.com/events/fatal06.htm
- 2007 - http://www.airsafe.com/events/fatal07.htm
- 2008 - http://www.airsafe.com/plane-crash/review-2008.htm
- 2009 - http://www.airsafe.com/plane-crash/review-2009.htm
- 2010 - http://www.airsafe.com/plane-crash/review-2010.htm
- 2011 - http://www.airsafe.com/plane-crash/review-2011.htm
- 2012 - http://www.airsafe.com/plane-crash/review-2012.htm
- 2013 - http://www.airsafe.com/plane-crash/review-2013.htm
- 2014 - http://www.airsafe.com/plane-crash/review-2014.htm
- 2015 - http://www.airsafe.com/plane-crash/review-2015.htm

During the study period, there were a total 162 significant events, with the first occurring on 9 July 2006, and the last on 4 November 2015. Note that this number may change after the date of this study's publication because new events may be added as information about past aviation-related events become available.

A note on the choice of sample size The choice of comparing a particular day's number of sessions to the distriibution of sessions for a recent 28-day period was used for two reasons:

1. Normal traffic varies during the course of a week, and may fluctuate for several reasons, including shortened work weeks, and events that attract unusual amounts of traffic.
2. Past observations of site behavior made it clear that extraordinary high traffic spikes were both rare and short-lived, with the effects of an event that sharply increases traffic typically dissipating after a week or less. Comparing a particular day's worth of traffic with distirbution of traffic over four weeks would make it less likely that the current day's number of sessions would not be compared against a distribution of traffic dominated by days of unusually high or low traffic.

Data preparation

The data representing the dates of a significant event were already process and available at the URLs listed in the previous section. The session data was exported from the 'Audience>Overview' section of the online application Google Analytics in the form of a CSV file (designated as sessions.csv) with several components:

- A six-line header that included the name of the web site and the dates covered
- Rows of data associated with each day, consisting of two columns:
  - Day Index: The date in a m/d/yyyy format
  - Sessions: The number of sessions
  - A final row with a single value in the 'Sessions' column that is equal to the sum of the sessions on the dates covered.

After the CSV file was downloaded from Google Analytics, the six header rows and the final were removed. Also removed were any of the initial rows of session data with a value of zero. Tracking of sessions did not begin until April 6, 2006, so all rows before that date were also removed. The resulting file was then loaded into R. That same data file was also made available for other researchers and is located at http://www.airsafe.com/analyze/sessions.csv.

```
# 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 meas
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 giv

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, starti
#        with the 22nd 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.

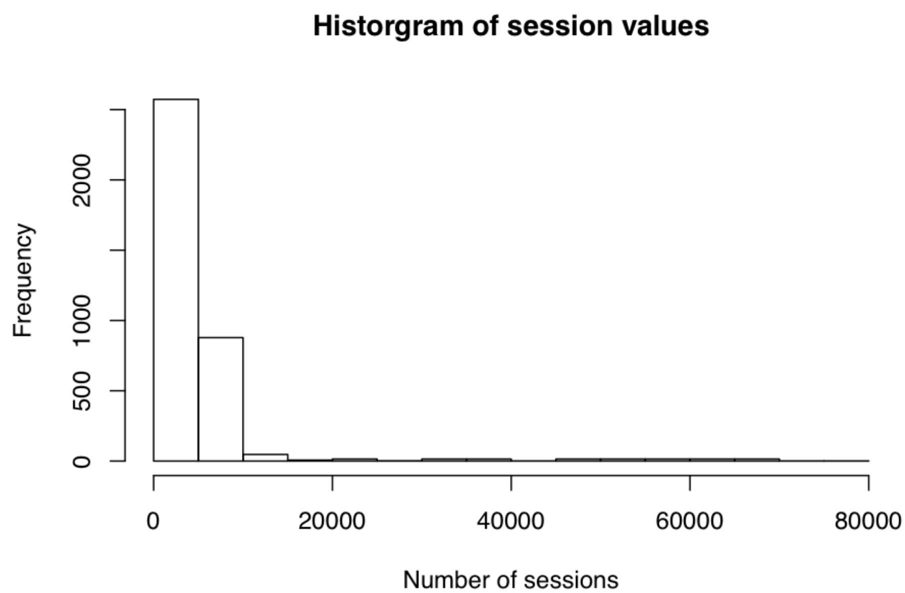
```
# Summary and histogram of the sessions data

summary(sessions$Sessions)
```

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

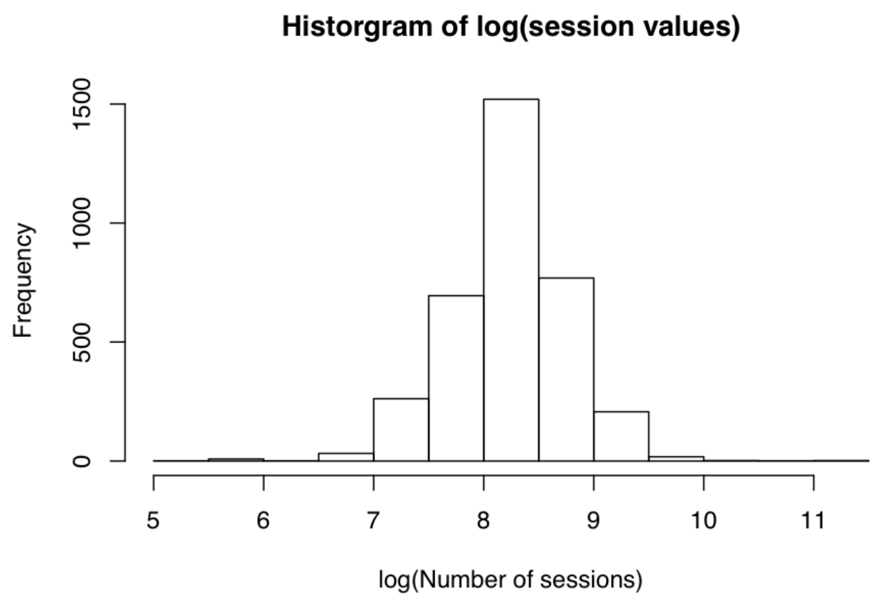
```r
summary(log(sessions$Sessions))
```

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

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




```r
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

```
# 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)
```

```r
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

The analysis was split into two parts. The first part consisted of comparing the dates with significant spikes in the number of sessions with dates of significant events on AirSafe.com to see which session spikes appeared to be diretly or indirectly associated with either significant events identified by AirSafe.com, or with other aerospace-related events. Because the date of a significant event on AirSafe.com is based on the local time when the event occurred, and because the Google Analytics setup is based on the Pacific Time Zone (GMT -7 during daylight savings time and GMT -8 otherwise), and because it sometimes take a number of hours before news media organizations becomes aware of the event, the first session spike associated with an event may occur the same day as the event, on the previous day, or one or more days after the actual occurrence of the event.

The second part of the analysis consisted of looking at those dates that had a significant session spike that could not be associated, either directly or indirectly, with an AirSafe.com-identified significant event. In these cases, by reviewing site traffic for that day, one could possibly find individual pages with unusually large traffic volumes, and through the use of tools like Google's search engine, uncover possible aeorspace-related events that could explain the high number of sessions.

Data overview of spike-related days The initial exploratory data analysis of the session data showed a wide range of values for the number of sessions for a spike day, from 1,638 to 75,120, with a median of 5808 and a mean of 7429. This reflects in part from both the steady increase in overall site traffic over the nine plus years of the study and the fact that some events generated an amount of sessions that were over an order of magnitude higher than most spike days.

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.

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.

```r
# 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
```

```r
# 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
```

```r
# 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
```

```r
# Chi-square test for the null hypothesis of a uniform distribution of spike days among the days of the
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
```

Identifying connections between spike days and significant events A combination of Google Analytics and Googles search engine were used to connect spike days with a particular event. Google Analytics provides an overview of traffic on any page on the site that carries the specific tracking code, so by using a feature of Google Analytics that provided details on page traffic on a pariticular range of days, it is possible to infer which pages were responisble for the spike in traffic.

When the page was specifically related to a significant event, the assumption was that the event, or something related to the event, led to a traffic spike. In the cases where the page was not specific to an event, for example, a spike on a page related to crashes of a particular aircraft model, the Google search engine was used to identify a likely cause of the spike. The default search engine setting is to search the entire web for online content that matches one or more keywords. Two optional settings were employed, either singly or in combination, in this study, limiting the search to a particular date range and limiting the search to news related items.

Double counting When two or more identifiable aviation-related events occurred on the same day as a spike, the events of that day are scrutinized to see if one of the following were true:

- At least one of the events would have been able to independently generate a spike day,
- Each event would have been able to independently generate a spike day, or
- No event have would been able to independently generate a spike day.

There were six dates where two aviation-related may have led to a spike event, and in two cases one of the two events led to a spike, in three cases both could have independently created a spike, and in one case, neither would have independently created a spike:

- 10 July 2006: A PIA F27 crashed on that day, and the day before a Sibir A310 also had a crash (spike for Sibir)
- 28 July 2010: Crash of a C-17 in Alaska, and an Airblue 321 in Islamabad (spike for Airblue)
- 20 April 2012: Crash of a Bhoja Airlines 737 the day after a hypoxia-related private aircraft crash that attracted substanial media attention (spike for neither)

- 23 July 2014: The crash of a TransAsia ATR 72 happened less than a week after the 17 July 2014 loss of Malaysia Airlines MH17 (spike for both)
- 24 July 2014: The crash of a Air Algerie MD83 the day after the crash of a TransAsia ATR 72 (spike for both)
- July 30, 2015: There was a significant increase in media attention around Malaysia Airlines flight MH370 after a flaperon was discovered on Reunion Island, and also around the television show "The Astronaut's Wives Club" (spike for both)

Spike events and spike days Becuase of the three days with two independent spike events, the remainder of this analysis will make a distiction between spike days (266 total) which is a day with a traffic spike, and with spike events which are spikes associated with a particular aviation-related event (269 total).

Comparing spike events to significant events During the period where this study was able to generate spike measurements (10 May 2006 to 14 November 2015), there were 162 significant aviation related events noted by AirSafe.com. Of those, 47 of the 162 significant events that occurred during the study period (29.0%) were associated with 141 spike events that occurred on or close to the day of the significant event. Three of the 47 significant events had an additional 10 spike events that were directly or indirectly related to a significnat event that occurred during the study period, but that occurred months or even years after the initial event. In total, 151 of the 269 spike events (56.1%) were due to significant events that occurred during the time period covered by this study.

An additional 17 spike events were associated with significant events that occurred before the study period, so 168 of the 269 spike events (62.5%) were associated with an AirSafe.com-identified significant event.

An additional 17 spike events were associated with an aviation- or aerospace-related event, so a total of 185 of the 269 spike events (68.8%) were related in some way to aviation or aerospace, implying that 84 spike events (32.2%) could not be directly or indirectly associated with aviation or aerospace.

Site traffic and celebrities Celebrities were associated with 21 of the 269 spike events (7.8%), yet only two of those spikes were due to significant events that occurred during the time period coverd by this study. Those two spikes were associated with the 9 December 2012 crash of the private jet carrying eight passengers and crew members, including singer Jenni Rivera.

The other celebrity-related spike events were directly or indirectly related to three events that preceded the study period:

- Dick Ebersol: Dick Ebersol and his two sons were passengers on a private jet that crashed in 2004, killing two of the three crew members and one of Ebersol's sons. From 2006 to 2015, the page associated with this crash was associated with a spike evet on 10 occassions. These spikes all appeared to be due to media attention around the perosnal arelationship between Britney Spears and Charlie Ebersol, the surviving son from the plane crash.
- Sandra Bullock: She escaped injury in the crash of a private jet in 2000, and the page associated with that crash led to a spike event in 2010 around news of a failed relationship, and another spike event in 2013 around media attention related to her movie 'Gravity.'
- Payne Stewart: This professional golfer, along with two pilots and three other passengers, died in a plane crash in 1999. The crash was due in part to the effects of hypoxia. During this study, a page on the site related to the Payne Stewart crash contributed to traffic increases that resulted in five spike events from 2009 to 2014, typically around anniversaries of his death, or during aviation events that were related to hypoxia.

Two spike events belong in a unique category. The 2015 television show ‘The Astronaut Wives Club,’ which was cancelled after its first season, focused on the wives several of the 1960s era NASA astronauts, and publicity about the show led to spike events centered on a page on AirSafe.com about deaths related with the US space program, and several of the astronauts mentioned during the show were also mentioned on the page. While the astronaut deaths occurred well before the study period, the public attention about these events was likely due only to the contemporary marketing efforts around the television series.

Top 10 spike events The top 10 days with the spikes with the greatest magnitude, specifically those days with a number of sessions with the largest number of standard deviations above the mean value of the number of sessions in the comparison period were as follows:

```
# The top 10 days with the spikes with the greatest magnitude, specifically those days with the largest

# Reording by spike
# head(spike.days[order(spike.days$SpikeSD, decreasing=TRUE),], n=5)
spike.reordered = spike.days[order(spike.days$SpikeSD, decreasing=TRUE),]


# Just the top 10
print("Top 10 spikes")
```

```
## [1] "Top 10 spikes"
```

```
head(spike.reordered[,c("Date", "Sessions","mean_range","SpikeSD", "Day")], n = 10)
```

```
##            Date Sessions mean_range  SpikeSD       Day
## 1154 2009-06-02    70251   3844.821 74.97874   Tuesday
## 3275 2015-03-24    75116   8967.500 60.88198   Tuesday
## 2649 2013-07-06    14351   2412.821 42.31466  Saturday
## 3276 2015-03-25    41278   8968.536 29.75973 Wednesday
## 1057 2009-02-25    28117   3502.643 28.00634 Wednesday
## 2894 2014-03-08    17305   1578.643 24.79717  Saturday
## 2650 2013-07-07     7921   2406.429 19.20985    Sunday
## 1016 2009-01-15    12212   2688.536 18.08898  Thursday
## 2877 2014-02-19     4639   1411.000 17.02753 Wednesday
## 1153 2009-06-01    17631   3833.571 15.59722    Monday
```

1. 2 June 2009 (74.98 SD) - Day after the loss of an Air France A330 over the Atlantic Ocean
2. 24 March 2015 (60.88 SD) - Crash of a Germanwings A320 in France
3. 6 July 2013 (43.31 SD) - Crash of an Asiana 777 in San Francisco, CA
4. 25 March 2015 (29.76 SD) - Day after the crash of a Germanwings A320 in France
5. 25 February 2009 (28.0 SD) - Crash of a Turkish Airlines 737 in Amsterdam, Netherlands
6. 8 March 2014 (24.79 SD) - Loss of Malaysia Airlines flight MH370
7. 7 July 2013 (19.21 SD) - Day after the crash of an Asiana 777 in San Francisco, CA
8. 15 January 2009 (18.09 SD) - Ditching of US Airways A320 in New York (“The Miricale on the Hudson”)
9. 19 February 2014 (17.03 SD) - Day after a Cathay Pacific 747 had a severe turbulence event over Japan
10. 1 June 2009 (15.6 SD) - Loss of an Air France A330 over the Atlantic Ocean

Findings

The analysis revealed that fewer than a third of the identified significant events during the study period (47 of 162 or 29.0%) were associated with a spike in the number of sessions, they were associated with 56.1% of all the spike events during the study period.

By using a combination of web site analysis using Google Analytics and the search options on the Google search engine, over two thirds of the spike events observed during the study period (185 of 269 or 68.8%) could be directly or indirectly associated either with one or more specific pages on AirSafe.com, or with another aviation-related event.

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_spikes.pdf
- Full analysis (Rmd) - http://www.airsafe.com/analyze/traffic_spikes.Rmd
- Full analysis (HTML) - http://www.airsafe.com/analyze/traffic_spikes.html