ASA 2009 data expo

Hadley Wickham

March 21, 2011

The ASA Statistical Computing and Graphics Data Expo is a biannual data exploration challenge. Participants are challenged to provide a graphical summary of important features of the data. The task is intentionally vague to allow different entries to focus on different aspects of the data, giving the participants maximum freedom to apply their skills. The 2009 data expo consisted of flight arrival and departure details for all commercial flights on major carriers within the USA, from October 1987 to April 2008. This is a large dataset: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed. The complete dataset and challenge are available on the competition website `http://stat-computing.org/dataexpo/2009/`.

Because the data was so large, we also provided participants introductions to useful tools for dealing with this scale of data: linux command line tools, including sort, awk and cut, and sqlite, a simple SQL database. Additionally, we provided pointers to supplemental data on airport locations, airline carrier codes, individual plane information, and weather.

1 Results

Nine groups formally entered the competition by presenting posters at the 2009 JSM:

- **First place**. Congestion in the sky: Visualising domestic airline traffic with SAS. Rick Wicklin and Robert Allison, SAS Institute.
- **Second place**. Delayed, Cancelled, On-Time, Boarding ... Flying in the USA. Heike Hofmann, Di Cook, Chris Kielion, Barret Schloerke, Jon Hobbs, Adam Loy, Lawrence Mosley, David Rockoff, Yuanyuan Huang, Danielle Wrolstad and Tengfei Yin, Iowa State University
- **Third place**. A tale of two airports: An exploration of flight traffic at SFO and OAK. Charlotte Wickham, UC Berkeley.
- **Honourable mention**. Minimizing the Probability of Experiencing a Flight Delay. Tanujit Dey, David Phillips and Patrick Steele, College of William & Mary.
- The Airline Data Set... What's the big deal? Michael Kane and Jay Emerson, Yale.
- Make a Smart Choice on Booking Your Flight! Yu-Hsiang Sun, Case Western Reserve University
- Airline Data for Raleigh-Durham International. Michael T. Crotty, SAS Institute Inc.
- Kaleidoscope Graphs. Mario A. Morales, Hunter College
- What Airlines Would You Avoid for Your Next Flight? Haolai Jiang and Jung-Chao Wang, Western Michigan University

In this special feature, the four winning groups present summaries of their findings. You can see the complete entries in poster form at `http://stat-computing.org/dataexpo/2009/posters/`.

2 Overview

Before we let you read the four interesting analyses that our finalists produced, here's a little context for the data. Figure 1 shows total number of flights per week. You can see a steady increase in flight numbers in the last 20 years, the striking effect of 9/11, and strong seasonal effects.

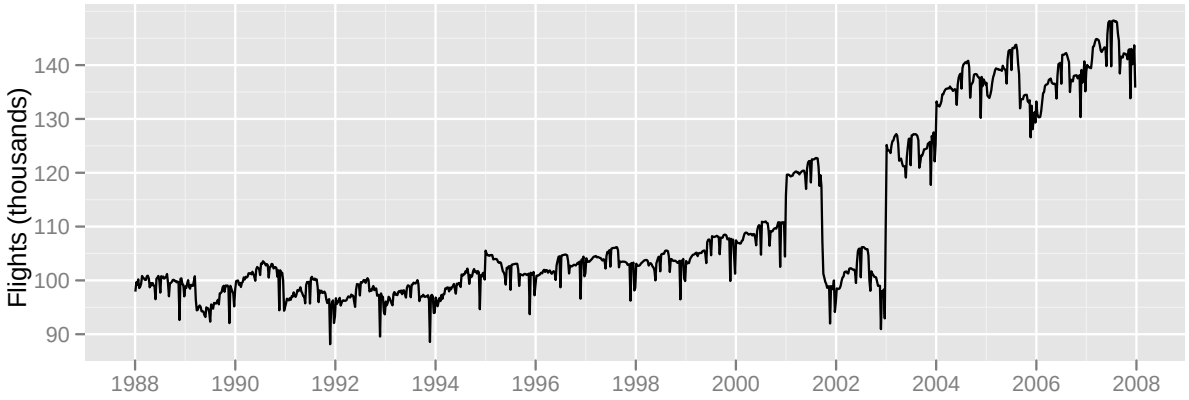

Figure 1: Number of flights per week from 1988 to 2008.

Figure 2 provides a little context as to what goes on within a week, showing for each year, the relative number of flights on each day of the week, relative to the average number of flights per day for that year. Weekdays have a fairly similar distribution, but different holidays have remarkably different impacts.

We'll leave the rest of the stories to the entrants.

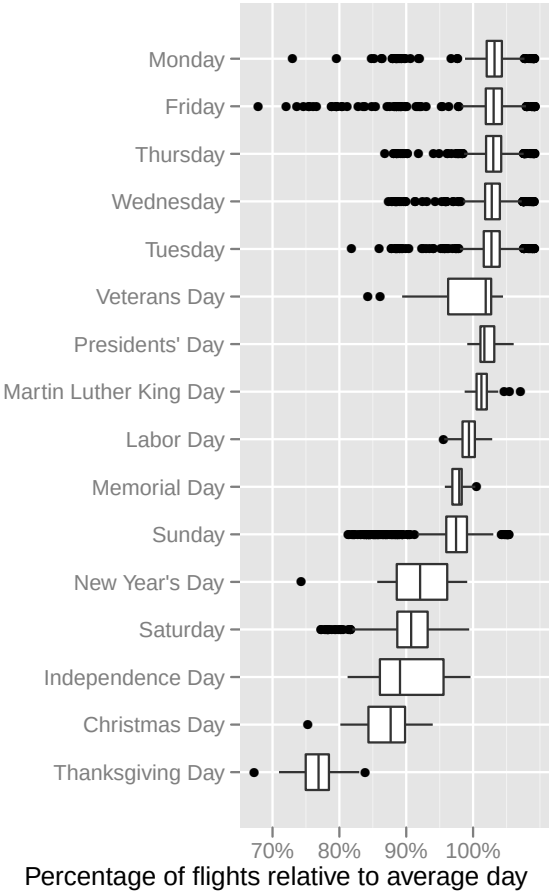

Figure 2: Within a week, the patterns are more complicated driven by day off the week, as well holidays.