Stat 222 Project 2: Visualizing Airline Data

1 Data Description

The data for this project comes from the U.S. Bureau of Transportation Statistics and their monthly "Air Travel Consumer Report." You will work with flight departure and arrival information for domestic commercial flights from 1987-2008, as well as some supplemental data on airports and airlines. The dataset is large, with over 120 million flights, and about 30 variables for each flight. We have created an SQL database for you to interact with the data; login details will be given in lecture.

The variables are described at `http://www.transtats.bts.gov/Fields.asp?Table_ID=236`. Because it's not well-described in the link, I'll also let you know that CRS stands for "computer reservation system", so that `CRSDepTime` and `CRSArrTime` are the scheduled departure and arrival times. All times are reported in the local time zone.

2 Your Assignment

Each group is responsible for creating a poster that visually answers a set of questions about this data. *Which* questions you answer is up to you, but think about telling a story. The story will be more interesting if the questions you address are related to each other in some way. I think you'll also find the story is more interesting if your questions involve two or more variables. For example, it would be interesting to know the distribution of flight departure times, but it would be even more interesting to know if this distribution changes with geographic location.

Note: this project was drawn from a visualization competition called the "Data Expo," sponsored but the American Statistical Association Sections on Statistical Computing and Statistical Graphics. You can find more details here: `http://stat-computing.org/dataexpo/2009/`. You are welcome to look at the entries to the competition for inspiration, although you should not simply replicate their plots. (You might also want to keep an eye out for Data Expo 2015!)

In terms of number of plots in your poster, a *minimum* number to aim for is 10, although it might be natural for some plots to be grouped together into a single figure, e.g. the same time series plot repeated multiple times for different airlines.

3 Initial Guidance (to do for first data debrief)

First, take a look at what tables are contained in the database, and what columns are present in each table. For each variable, consider what you expect to see, and then devise an SQL query and perhaps some `R` code to check whether your expectations are met. You might want to split up the variables across your team members. Make a note of anything unusual that you find.

Next, working individually, brainstorm *three* questions to use a starting point for exploring this data. The answers to these questions will almost certainly suggest other questions to you, but use these as a starting point. For each question,

- Think about what plot you could make to help answer the question.
- Describe exactly what data or data summaries you need to make the plot.
- Write an SQL query to obtain the data.
- Make an initial version of the plot. (You may want to refine it later.)
- Interpret the plot. Does it suggest any other questions/plots to consider?

These steps may sound obvious, but I know from experience that the temptation is there to jump in and start writing code before you've given much thought to what you want to learn. I think you'll find you get richer and more interesting results in an open-ended project like this if you allocate more of your time at this stage to *thinking* and less to implementation.

As you work, it would be useful if you use `system.time` to record the time it takes to run each query from within `R`. We can identify things that are running slowly and try to optimize the database to give you better performance.

4 Next Steps

On Wednesday you will discuss your findings with your group and decide on an overall theme for your poster, as well as additional plots that you want to make.

One thing I'll also ask you to consider on Wednesday is whether there are any additional datasets that you want to obtain. For example, you could look at population data for the cities corresponding to each airport, or perhaps weather data. What about financial data for the various airlines?

5 Poster Details

There are many ways to create a poster. Here are some possibilities for you to explore:

- Beamer Poster Package
  `http://www-i6.informatik.rwth-aachen.de/~dreuw/latexbeamerposter.php`
- Powerpoint (try searching for "research poster powerpoint template")
- Adobe Illustrator or InDesign (available free to students at `https://software.berkeley.edu/adobe`)
- Pages (recommended over Keynote for posters)

Your posters will be due *at the latest* on Tuesday, February 24, and we'll spend some time sharing them in class on Wednesday, Februrary 25. (You should have everything finished by Monday, so that you have time for printing.) A service I can recommend is `http://gif.berkeley.edu/services/printing.html`. Note that they require an appointment, which I recommend you go ahead and schedule now.

If you use this printing service, your poster should be 36" or 42" along one side. (These are the paper sizes they stock.) 36" high x 48" wide is fairly common poster size.

When designing your poster, you may find it helpful to draw the layout before you try to implement it in software. Another piece of advice is to be careful with the resolution of your figures. After designing your poster, you may need to regenerate them at the actual size they will occupy. Taking a small figure and enlarging it to fit the poster will cause the image quality to be poor.

The poster should contain a title, your names, data source(s), the plots, and text to tie everything together. Someone should be able to understand the main findings simply by reading it, but be careful not to include so much text that the poster becomes difficult to skim. It's typical to "present" the poster and give more details verbally. The Data Expo poster entries are good models to emulate.