Project Description:

Lrm96, ss2644, jlp277

Description of Data:
We got our data from DataExpo 2009 and used data from the year 2008:
http://stat-computing.org/dataexpo/2009/the-data.html
http://stat-computing.org/dataexpo/2009/supplemental-data.html
For a full explanation of values, use this link:
http://www.transtats.bts.gov/Fields.asp?Table_ID=236
We also used d3 libraries such as geo and linear.
For inspiration we primarily used examples in class, though we also used some examples by mbostock: http://bl.ocks.org/mbostock/7608400

The data consists primarily of data about airplane flights between airports around the world. For our visualizations, we focused only on airports and flights that took place in the United States.

For each flight there were originally 29 pieces of data, but we decreased the flight data to only include the origin of the flight (IATA code as a string), the destination of the flight (IATA code as a string), the month the flight took place (integer 0<x<13), delay after arrival (integer), delay at the departure site (integer), and whether or not the flight was cancelled (0 if not cancelled, 1 if cancelled). This data was to show how delays would take place relative to variables such as location and time of year. This data was contained in files data.json (large file only for the top 18 USA airports, which also included data points for flight number and cancellation code which we found were not significant), and in smalldata.json which was a small sample which uses all ports. We needed to decrease the amount of data in these files, and the method we used was merely taking every i-th element of the original data. This method was selected because the data was organized by month and airport. Taking elements mod i was the easiest way to get data from various months and airports while not having bias towards certain airports.

For the airport data in the supplemental data, we only kept the IATA code (string), the latitude of the airport (number), and the longitude of the airport (number). This data was used only for placing airports onto the map. This data is contained in ports.json (only has data for the top 18 airports) and allports.json (which has data for almost all airports in the USA).

We also used data from us.json and usda-atlas-people.csv that was provided in class to create a map of the USA that allows for elements to be placed on the map according to latitude/longitude coordinates. Colorbrewer was also used from http://colorbrewer2.org/ and data from colorbrewer can be found in colorbrewer.js .

Delay.csv was created from the entire 2008 flight data-set. For each month, we aggregated the number of Cancellations, DepDelay15 (which is the number of flights departed more than 15 minutes late) and ArrDelay15 (which is the number of flights arriving more than 15 minutes late) fields from the data-set and stored the result in delay.csv.

Description of mapping of data to visual elements:

For the first two visualizations, the primary variables were location and color. Data from each airport was collected from smalldata.json and the average delay of each airport (as origin) was found. As long as an airport would occur on the map of the US and there was at least one piece of data about the airport's delays. Location was determined based on latitude and longitude, and was placed on map through the use of d3 libraries. Color was determined through the use of Colorbrewer and d3's linear scale method. Shape was held constant by making all of the airports circles. When it was found that using all airports would result in a very crowded visualization, we decided to also make a version that was not so crowded and only used the top 18 airports in the USA (the ones people would be most likely to care about).

The second visual is of two undirected networks of airports with edges representing flights to and from either airport to the other. Widths of edges indicate volume of flights and colors indicate aggregate delay. Routes with more delay are redder, and routes with less delay are greener. Additionally, to reduce clutter, routes with insignificant amount of traffic were omitted (the threshold for omission was determined by aesthetics).

Total delay of a route is the sum of delay of all flights on that route where delay for each flight is measured as minutes late/early at destination gate. Delays are then averaged over number of flights and normalized to a [0, 100] scale for coloring purposes. Color is determined using the normalized delay scores and applying them to a continuous linear scale from green to red. This system of coloring is useful as most delay scores of the routes are fairly evenly distributed from 0 ~ 40 and a few are sparsely distributed from 40 ~ 100.

Unlike the first two visualizations, the third visualization focuses on how cancellations and delays vary across months of the year 2008. This visualization aggregates data from all the regions in the US. Data for this visualization comes from delay.csv. We represent the information in a single multi-color bar chart since the data is discrete. Inspiration for this visualization came from mbostock github gallery. We represented all the data in a single bar chart without compromising on comprehensibility. All information related to a particular month was plotted in a single multi-color bar. Different colors were used to represent Cancellations, Arrival and Departure delays.

Story of the Data:

Originally, we focused on the idea that airports in certain regions may tend to have more delay versus airports in other regions. We used the first visualization to get an idea of how tends may occur in the USA overall. We realized that looking at all the data for the USA would not be feasible--we needed to reduce the number of airports we were examining. We tried using the top 18 airports of the USA (selected on the basis of the top 25 airports in the USA, and then eliminating ones that were too close together and would have caused crowding). Unfortunately, this did not reveal much since all 18 were a little late or close to on time.

We explore the different paths between the data to figure out which routes tend to be faster in general between the airports. Looking between the maps that show routes, it seems that the West Coast and Northern East coast have a tendency to be relatively on time or early while almost all flights related to O'Hare (ORD) tend to be late. These results can be seen faintly on the previous visualization via clusters of on time and early dots on the West Coast and parts of the East Coast. This information is useful for choosing which airports to frequent, but there is still the question of when is the best time to fly.

We used all data available to us created a histogram that shows the number of flights that departed over 15 minutes late and arrived over 15 minutes late, as well as the number of cancelled flights. This data confirms many popular beliefs: that more flights are cancelled in the winter months of January and February, and more flights take place in the summer months. However, what might be surprising is that flights never seem change the amount that they are late. If they leave later than 15 minutes, they always arrive 15 minutes late. This implies that weather and time of year are accurately represented by the airlines' estimates and if you depart late there is little to no hope that you will arrive on time.

In short, the best places to fly out of are the West Coast and the Northern East coast, as well as any number of small airports in the Rocky Mountain region. Avoid airports like Chicago O'Hare and the San Francisco International Airport if you are flying domestically. Flying in the winter does increase the chance that your flight will be canceled, and, unfortunately, if you leave late you can basically give up on your chances of arriving on time.