INFO 3300
Travis Beach (trb96), Zichuan Zhou (zz278), Candy Lin (cl839)
April 23, 2015
Project 2

The Data
        The data for this project came from several sources. The population statistics come from US census data (http://www.census.gov/compendia/statab/2012/tables/12s0020.pdf, with 2014 data coming from estimates). We were able to export as an excel(.xls) file. The data included both metropolitan areas as well as the micropolitan statistical areas they are comprised of. We were able to filter out the smaller regions pretty easily. The fifty largest areas were chosen based on their 2014 populations, and data from the previous years was entered based on that. (Because the ranking isn't the same for each year, we chose to pick a year, and rank them relative to that.)
        The flight data was a little trickier. The data was from the Bureau of Transportation Statistics (http://www.transstats.bts.gov/DL_SelectFields.asp?Table_ID=258&DB_Short_Name=Air%20Carriers). There, we found a tool that allowed us to select which fields we wanted for each market. We first had to consolidate based on airline. Also, because our visualization concerns metro regions rather than the airports themselves (ex. New York Metro is served by JFK, EWR, and LGA), we had to consolidate some of the airports in a single category. We were able to write a python script (filter.py) to accomplish both of these simultaneously. CSV files of the curated data were created and imported.
The methodology of how we include airports is to use the DOT's hub definition (30 large hubs and 33 median hubs). If the city has airport(s) on the hub list, we include all of them and exclude all other auxiliary airports (for example, we exclude Stewart Airport for NYC). If the city doesn't have any airport on the hub list, we choose its major airport no matter how small the airport is.
        Based on these data, we are able to accomplish the visualization and the information side bar for each city and city pairs.

Scales
        Circles: the circles are used to demonstrate the size of the population. Their radius is determined by the square root of the population, to ensure that the relationship between the areas of the circles is preserved. A key is provided to show the circles from a variety of sizes.
        Lines: The lines are colored based on the passenger traffic on that route. Routes with the heaviest traffic are the darkest, with lighter colored lines representing less popular routes.

The Story

The purpose of this visualization is two-fold. First, it seeks to examine the relationship between population and flight patterns. Naturally, cities like New York and Chicago support a lot of traffic, and well-known hubs like also Atlanta have a lot of flights relative to the size of the city, but other interesting data points also emerge. Orlando, a center for tourism, has a large amount of flights despite maintaining a small metropolitan population. It is also interesting to look at a city like Cleveland, and see how the flights has decreased as the city has depressed over time. Chicago, as the country's second hub after New York in terms of total passenger traffic, is the only city that is connected to all 50 metropolitan areas. New York is second to Chicago in terms of destinations.

The second purpose is to look at how a city pairs are related. Interesting relationships develop between large metros that are relatively close. For example, one can drive from New York City to Washington D.C. In about four hours, but there are still more than 2,000,000 flights between the cities each way annually. This likely indicates that flyers are there on business, rather than tourism or visiting family. Professionals would likely need to get back and forth between the two cities faster than other modes of transportation allow, while other travelers would opt for the cheaper, but slower route. In addition, based on our analysis, the busiest route in U.S. is New York – Miami route. Note that this contradicts Wikipedia's data, which indicate New York – Chicago as the busiest, because we might have included different airports (or Wikipedia is wrong). The whole busiest route list is very similar, though.