1. **New York Times** (50 points, Adapted from Rachel Schutt at Columbia) There are 31 datasets named nyt1.csv, nyt2.csv,...,nyt31.csv, which can be found on my webpage. Each one represents one day's worth of ad impressions and clicks on the New York Times homepage in May, 2012 (these are simulated). Each row represents a single user. There are 5 columns: age, gender (0=female, 1=male), number of impressions (page views), number clicks (actions) and whether the user was logged.in.

WARNING: These data are designed to simulate real data (and all the accompanying headaches). Don't make assumptions (e.g., that age > 0) unless you've verified that the data actually comply with these assumptions.

a. Create a new variable, age_group, that categorizes users as "<18", "18-24", "25-34", "35-44", "45-54", "55-64" and "65+".

b. For a single day,

   1. Plot the distributions of number impressions and click-through-rate (CTR=# clicks/# impressions), for these 6 age categories. [You will turn in a plot called "single_day.pdf" with these results]
   2. Define a new variable to segment or categorize users based on their click behavior
   3. Explore the data and make visual and quantitative comparisons across user segments/ demographics (<18 year old male vs < 18 year old females or logged-in vs not, for example). ["comparison.pdf"]
   4. Create metrics/measurements/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, mean, median, variance, max, and these can be calculated across the various user segments. Be selective. Think about what will be important to track over time; what will compress the data, but still capture user behavior. [choose one and save an indicative plot as "metric.pdf"]
      i. Now extend your analysis across days. Visualize metrics and distributions over time. Your plot should emphasize what actually changes over days. ["time.pdf"]
      ii. Describe and interpret any patterns you find. [Include in your writeup]