W4242 Intro to Data Science
Schutt

Homework #1
Due: Wednesday, September 19, 2012

(1) New York Times:
There are 31 datasets named nyt1.csv, nyt2.csv,...,nyt31.csv, which can be found here: http://stat.columbia.edu/~rachel/datasets/nyt1.csv [This is the path to the first one. Change the # accordingly]
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 impressions, number clicks and logged.in.

You can load a single file into R with this command:
data1 <- read.csv(url("http://stat.columbia.edu/~rachel/datasets/nyt1.csv"))

(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,
-- Plot the distributions of number impressions and click-through-rate (CTR=# clicks/# impressions), for these 6 age categories.
-- Define a new variable to segment or categorize users based on their click behavior
-- 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).
-- 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.

(c) Now extend your analysis across days. Visualize metrics and distributions over time.

(d) Describe and interpret any patterns you find.

(2) Case study:
You have been hired as Chief Data Scientist at realdirect.com, and report directly to the CEO. They (hypothetically) do not yet have their data plan in place. They're looking to you to come up with their data strategy. Here are a couple ways you should begin to approach this problem.

(a) Explore their existing website, thinking about how buyers and sellers would navigate through it; and how the website is structured/organized. Try to understand the existing business model, and think about how analysis of realdirect user behavior data could be used to inform decision-making and product development :
-- Come up with a list of research questions you think could be answered by data
-- What data would you advise the engineers log and what would your ideal data sets look like?
-- How would data be used for reporting and monitoring product usage
-- How would data be built back into the product/website
(b) As there is no data yet for you to analyze (typical in a start-up when they are still building their product), you should get some auxiliary data to help gain intuition about this market. Go to: http://queens.about.com/od/realestateandapartments/a/recent-sales-ny.htm
Click on: Rolling Sales Update (after the 5th paragraph). You can use any or all of the data sets here. Start with Manhattan August, 2011-July 2012.

-- First challenge. Load in and clean up the data. Compared to the first problem, which had clean simulated data, this data will be messier. If you are new to R, this process can seem daunting and time-consuming. This is where Jared comes in. He will help with these steps at the first lab on Monday, September 10. However you should try on your own first. Once you get it loaded in, conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, making sure values you think are numerical are being treated as such, etc.

-- Once the data is in good shape. Conduct exploratory data analysis to visualize and make comparisons (i) across neighborhoods, and (ii) across time.

-- If getting this far was enough work (especially if you are new to R), stop exploring the data. If you get into it, go further than the above, finding meaningful patterns in this data set.

For both (a) and (b), summarize your findings in a brief report aimed at the CEO.