Introduction to Data Science: CptS 483-06 – Fall 2015

Assignment 2: Exploratory Data Analysis

Release Date: 2015-09-11. Due Date: 2015-09-16.

There are 31 datasets named nyt1.csv, nyt2.csv, . . . , nyt31.csv, which you can find here: https://github.com/oreillymedia/doing_data_science. (The datasets are also made available in a zipped file under the folder Datasets on the course's webpage: http://www.eecs.wsu.edu/~assefaw/CptS483-06/Datasets/dds_ch2_nyt.zip.) Each one of these datasets represents one (simulated) day's worth of ads shown and clicks recorded on the *New York Times* page in May 2012. Each row represents a single user. There are five columns: age, gender (0=female, 1=male), number of impressions, number of clicks, and logged in.

You will use R in this exercise to handle these datasets. Load the data in the file nyt1.csv into R. You can do this using the following R command:

```
data1 <- read.csv(file = "nyt1.csv")
```

(assuming you have downloaded the file nyt1.csv into your working directory) or

```
data1 <- read.csv(url("http://stat.columbia.edu/~rachel/datasets/nyt1.csv"))
```

Here are the problems you will solve in this exercise.

1. Create a new variable, `age_group`, that categorizes users as "<18", "18–24", "25–40", "41–64", and "65+"
2. For a single day:
    - Plot the distributions of the number of impressions and click-through-rate (CTR = Nr. clicks/Nr. impressions) for these five age categories.
    - Define a new variable to categorize (segment) users based on their click behavior.
    - Explore the data and make visual and quantitative comparisons across user segments/demographics. In particular, compare under 18-year-old males versus under-18-year-old females, and 18 to 24 year old females versus 25 to 40 year old females.
    - Create metrics/measurements/statistics that summarize the data. Metrics you should include are CTR, quantiles, mean, median, variance and max, and these should be calculated across the various user segments. Think about what will be important to track over time—what will compress the data, but still capture user behavior.
3. Now extend your analysis across days. Visualize some metrics and distributions over time.
4. Describe and interpret any patterns you find.