**Homework 1: Exploratory Data Analysis**

**Goals:**

Use R to do exploratory data analysis including pattern finding.

**Description:**

Students are divided into groups of 4-5.

Students are given 31 datasets named nyt1.csv, nyt2.csv, ...nyt31.csv, which can be downloaded on WISE, under the homework 1 folder. Each file represents one (simulated) days worth of ads shown and clicks recorded on the New York Times homepage in May 2012. Each row represents a single user. There are five columns: age, gender, (0-female, 1-male), number impressions, number clicks, and logged-in.

Here are some suggested steps you may follow:

1. Manually open 1-3 data files and report your first impression of the data, you may describe anything you have observed.
2. Create a new variable, age\_group, that categorizes users as “<18”, “18-24”, “25-34”, “35-44”, “45-54”, “55-64”, and “65+”.
3. For a single day:
   1. Plot the distributions of number impressions and click-through-rate (CTR = # clicks/#impressions) for these six categories.
   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 segment/demographics (<18-year-old males versus < 18-year-old females or logged-in versus not, for example)
   4. Create metrics/ measurement/statistics that summarize the data. Examples of potential metrics include CTR, quantiles, mean, median, variance, and 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.
4. Extend your analysis across days. Visualize some metrics and distributions over time.
5. Describe and interpret any patterns you find.

You may refer to the textbook P39 for some sample code to get it started. Be creative with visualization and pattern finding.

**Deadline:** Feb 7 11:59PM.

**Deliverables and submission:** Each team will develop R code collaboratively and write a report to describe their findings. Each team will also need to make a powerpoint presentation to report their findings.

The team leaders need to upload the zipped code and written report to WISE dropbox before the deadline. During Feb 8 class, each team will give a 5-10 minute presentation about their findings.