MSiA 421
Professor Edward Malthouse

**Homework 1**

**Due: January 24, Start of class**

Submit your answers on Canvas. I will crate a "quiz."

You may use any software you like to complete this exercise. The data are available on Canvas under Files/hw.

1. Explore the data. Understand the variables you have. Run basic descriptions on the variables to confirm that they make sense. Make a list of issues you find.

2. Create variables as we discussed in class by aggregating the web data. See my examples using `dplyr` and `spread`. Submit a list of your variables, give basic descriptives (e.g., $n$, mean, min, max, sd).

3. Identify 5–12 clusters based on reading behaviors. Remember the ultimate goal is to use the clusters to personalize a newsletter to readers each day and perhaps develop other specialized newsletters/products, e.g., see SRC Wine competition or NYT Crosswords. Here are a few thoughts:

   - You will certainly want to consider the content tags, e.g., sports teams, wine, local news, etc.
   - You might want to consider device (phone vs. PC), referal source, time of day, etc.
   - High-level summarizes such as days/week reading, pages/session, time/page, breadth of reading could be used.
   - You might want to consider the crossed-basis approach, where you develop two separate cluster analyses and cross them.

   Submit your clusters with names. List which variables were used to define the clusters. Give the cluster sizes and means on the variables used for clustering and profiling. Discuss how you would recommend using the clusters.