**Working title: Who’s the chump?**

**Topic Question**: Who compromises more on television viewing preferences upon marriage—men or women?

**Data to be used**: 1) Television viewing data from individuals from 30,000 households across the US over a 5 month period (I’m calling this dataset: ‘main’) and 2) Demographic table with demographic characteristics of each household and each panelist (individual), called ‘demo.’

**What I know about the data**: Main will be quite large, somewhere between 1 and 2 gigs. It does not currently contain demographic information, which I’ll need, so demo dataset needs to be tracked down. .Variables needed included are age, gender, marital status, presence of children.

To get the main dataset to a manageable size I’ll eliminate data that does not relate to answering the main questions. This means dropping households that don’t qualify based on demographic composition or other factors. For example, I’ll have to screen for a balanced mix of men and women, both single and married across age ranges. I’ll eliminate households with kids, individuals under the age of 18, and perhaps households that don’t watch very much TV. I’ll also have to develop a way of screening for outliers, as I’d expect there will be some wacky records that could skew the results. To keep it manageable, I’ll also focus only on the top 100 shows. If further trimming is needed, I’ll focus on a few key geographies.

**Analysis plan**:

1. Merge demo with main, and create combo variables

* Create 5 new variables, and add them to main. These are age\_group, gender and mar\_status variables for each panelist. Create an age\_gender variable, and demo\_group variable that combines age, gender and marital status.
  + There are 6 age groups: 18 to 24 (inclusive), 25 to 34, 35 to 44, 45 to 54, 55 to 64 and 65+, and gender and marital status are binary, so there should be 12 age\_gender groups and 24 demo\_groups.
  + Consider adding geography for later use (not sure about this)

1. Classify TV programs:

* Starting with main, create ‘genderskew’ variable for each program (classify each of the 100 programs as ‘chick’, ‘dude or ‘neutral’)
  + For each month, for each program, sum the total viewing minutes for men, women and in total. Calculate % of minutes that are female.
    - Play with the data to determine the right parameters, but come up with a decision rule that categorizes each program into one of these three buckets based on the composition of its total viewing audience.
      * For example, for American Idol, suppose 60% of the audience is female; for Pawn Stars, 70% was male; and for Law and Order, 52% is male, 48% female. These programs would be categorized as ‘chick’, ‘dude’ and neutral, respectively.
    - Note that # of programs will be >100 across 5 months because top 100 are not constant, and programing changes from month to month (FIFA only on in the summer months, for example) Also note that it’s likely after the first pass that some programs will have multiple genderskew assignments in different month. (Law and Order could skew women one month and men the next.) Each program to have one genderskew over the whole time period, so this might need some manual adjustment.
  + Result should be a program file containing 100+ records, one for each unique program over the 5 month period, with a ‘genderskew’ variable as chick, dude or neutral. Make sure each program has one and only one assignment.
* Merge genderskew variable into main based on program variable

1. For each program, determine likelihood of viewing for singles of each age/gender combination. This will to be used as ‘control’
   * Create lovS variable (likelihood of viewing for singles). For each program, in each month, calculate total viewing minutes in total and by age\_gender group when mar\_status = single. Lovs = (total viewing minutes for singles in that age\_gender group)/(total viewing minutes for that age\_gender group)
     + *If we’re trying to predict lov for a given 33 year old woman in a given month, we’d assume that = lovs for her age\_gender group.*
   * Create lovM variable (likelihood of viewing for marrieds). For each program in each month for each married, calculate total minutes
     + Note: lovS will have a value for both singles and marrieds, but for singles, lovM = NA
2. Compare married viewing behavior to singles within age\_gender groups
   * Create viewing behavior index varlable (vbi.) This is an index that reflects the extent to which married behavior is different from single behavior within age\_gender groups. (vbi = (lovM/lovS)\*100.
     + A vbi of 100 means for that individual for that program, being married made no difference to their viewing behavior. An index of >100 means marrieds were more likely to view the program; <100 means marrieds were less likely.
3. Determine if the difference in vbi is statistically different between single and married groups, and if so, how. If different, the group whose vbi changes more is officially the ‘chump’ group.
   * Try slicing by available demo variables to see if findings change. Is it true for some age groups but not others? Does it vary by day or time (weekends vs weekdays; daytime viewing vs. night?) Using some of the data as a holdout sample, build a model to predict viewing behavior based on these variables.

**Rationale for selecting this topic**: I chose it because I come from the digital world but want to get some experience with television data. Also, my husband and I thought I t would be fun to settle this marital debate! :)