PSY652, Unit 2, In class activity – Data Wrangling

In this activity you will build on the Fandango R NOTEBOOK.

Dataframe: In Dropbox Unit1: fandango.csv

Description of Variables: In Dropbox Unit1: Fandango Dataframe Variable Descriptions

1. Open the Fandango Notebook. At the bottom of the notebook (after all of your graphs), create a first level header and call it: A collection of wrangling techniques (be sure to put a single hashtag before this phrase).
2. Create a second level header called: filter examples. Then add a code chunk just below this header (INSERT > R). Create a subset dataframe called film\_2014 that includes all films from 2014. Now, modify this line of code to choose all films in 2014 for which the rottentomatoes score is greater than or equal to 80 and the rottentomatoes\_user score is greater than or equal to 80. Click on the datafile in the upper right quadrant (environment) to make sure it worked.
3. Create a second level header called: select examples. Add a code chunk. Create a dataframe that includes only these variables: film, rt\_norm\_round, rt\_user\_norm\_round, metacritic\_norm\_round, metacritic\_user\_norm\_round, imdb\_norm\_round, fandango\_stars. Call this dataframe normed (these are all of the normed rating scores – see the data description file). Click on the datafile in the upper right quadrant (environment) to make sure it worked.
4. Create a second level header called: mutate examples. Add a code chunk. Use mutate on the normed dataframe you just created to create a new variable (call it one\_to\_see) inside this dataframe that equals 1 if all 6 of these normed scores are greater than or equal to 4, otherwise a 0.
5. Create a second level header called: arrange examples. Add a code chunk. Use arrange on the normed dataframe that you just updated with the new variable called one\_to\_see. Sort the dataframe so that the films coded 1 on one\_to\_see appear at the top.
6. Create a second level header called: summarize examples. Add a code chunk. Using the normed dataframe, first group by (use group\_by) fandango\_stars and call this grouped dataframe normed\_stars, and then request the average normed rotten tomatoes score (rt\_norm\_round) for each level of fandango\_stars (call this summary variable avgscore).
7. Create a second level header called: pipe examples. Add a code chunk. Write a pipe to do the following (name the output of the pipe (my\_first\_pipe):
   1. Pull in the fandango dataframe
   2. Filter to include only 2015 movies
   3. Select these two variables: file, rt\_norm\_round
   4. Create two new variables, best (equals 1 if rt\_norm\_round equals 5, 0 otherwise) and worst (equals 1 if rt\_norm\_round = .5, 0 otherwise)
   5. Filter to include only the best and worst movies (i.e., a score of 1 on best or worst)
   6. Arrange by worst, so that the best movies are on the top.

