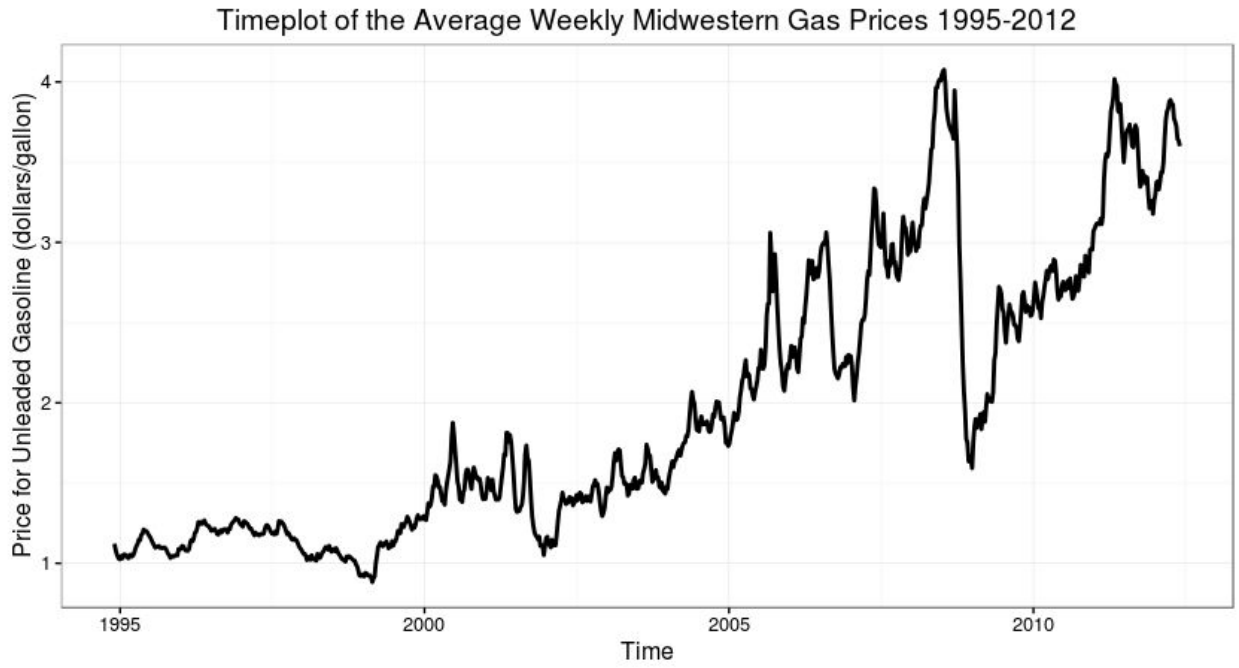
STA 404/504 Homework 2 - Data Cleaning

Submission through Canvas - Due Monday 2/18/19 by 5:00 PM

Learning Objectives:

- Forming a data cleaning plan, breaking into tasks
- Reading in data with poorly structured heading
- Restructuring specific columns from wide format to tall format
- Time formatting variables
- Documenting a fully reproducible data cleaning/visualization process in a scripted language

Assignment:
The data file *midwest_gas_prices.csv* is attached to the assignment link for Homework 2. This file contains the average weekly midwestern fuel prices for a gallon of regular unleaded gasoline from November 1994 to May 2012. We want to use this data to create a time series line-plot of the data that looks like the image below:



The problem is that the data is very messy and improperly structured to create this visualization. Your task is to clean and restructure the data ***in R*** using the *dplyr, tidyr, stringr,* and *lubridate* packages so that the cleaned data is saved as data frame that has only two columns: date and price. The date column should be converted from a character to a POSIX format using a function from the *lubridate* package. Open the data file in a spreadsheet editor like excel to take a look at the structural issues.

- The header of this data is two rows of poorly formatted labels. When you read in the data you will want to skip over the header using the options in the *read.csv(.)* function.
- The first column has the year/month combined followed by five pairs of columns for the dates and prices associated with weeks 1 through 5 of each month. Each of these five column pairs will need to be moved from wide format to tall format using functions in the *tidyr* package. Note that this will be made much easier if you first create separate data frames with the columns of dates and columns of prices, then gather the columns in each, then recombine after they have both been reformatted wide-to-tall.
- Also notice that the years for each date are listed as 2016, this needs to be replaced with the real year from the year/month column. I suggest using a combination of the *stringr* package and the paste() function to fix the year.
- After you get the date column looking like “mm/dd/yyyy” or “mm/dd/yy” you can use the *mdy(.)* from the *lubridate* package to transform the column to a POSIX date-formatted variable.

Submission Format: **R code file**
For this homework you will submit R code that can read in data from the original file “*midwest_gas_prices.csv*”, loads necessary packages, cleans the data and creates a plot as identical to the timeplot above as possible. Please be sure to properly document and organize your code so that the data cleaning and plot creation processes are clear.

Note: For this assignment the ***entire*** data cleaning process must be conducted in R, no cleaning “by hand” may be done in excel before loading to R. For this homework you will only be turning in the R code.

What will I look for in grading?
5 points - Successful data cleaning process (header, wide-to-tall, date formatting)
3 points - Clarity of documentation, structure, and comments in code
2 points - Similarity your timeplot to figure above (structure, labels)