STA 141A

Fall 2016

Homework 4

Due: November 29 (Tuesday) at 5:00 PM

Submit the assignment both electronically (through smartsite) and by submitting the printed copy. The electronic submission must be in the form of a zip folder (with extension .zip, .7z, etc.) containing two files: (i) descriptions of your analysis (as appropriate); (ii) codes used.

Honor Code: *"The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment:" (ADD: names of persons or web resources, if any, excluding the instructor, TAs, and materials posted on course website)*

1. The data in the zipped folder `CarAdvert.zip` consist of 1531 posts on advertisements of used vehicles by several vendors. The task is to extract relevant information as described below and create a data frame in R to represent the information. Finally, save the information in the form of a comma separated spreadsheet (.csv) file.

   (i) Model (Year and Make) of the car

   (ii) Vehicle Identification Number (VIN)

   (iii) Price

   (iv) Mileage

   (v) Color (Interior and Exterior)

   (vi) Transmission

   (vii) Engine displacement (in liters)

   (viii) Name of company selling the car

   (ix) Street address of the company

   (x) Phone number of the company

   (xi) Website of the company

   Some of these information may not be available for particular posts. Also, some names of models may be misspelled. Try to correct the latter wherever possible. You may need to perform a basic statistical summary of the Make of the cars in order to find the right spelling.

   You may find the following functions useful: readLines(), grep(), grepl(), agrep(), regexpr(), gregexpr(), regmatches(), sub(), gsub() and functions in the **stringr** package.

2. Use the **XML** package and text processing to extract the information from the webpage (list of publications by Professor Hans-Georg Müller):

   `http://anson.ucdavis.edu/~mueller/cveng13.html`

   Create an R data frame containing the following information for each of the journal publications:

   (i) Year of publication.

   (ii) Authors.

   (iii) Title of publication.

   (iv) Journal title.

   (v) Journal volume.

   (vi) URL for the publication.

   Do a basic statistical summary of the result (including information such as number of publications per year, number of co-authors, number of publications in different journals, etc.).