Advanced Data Analysis 2, Stat 428/528 Homework 18 Name: _____
Spring 2015 Prof. Erik B. Erhardt

Part I. (90 points) Do all calculations in R. All R code for the assignment should be included with the part of the problem it addresses (for code and output use a fixed-width font, such as Courier). Code is used to calculate result. Text is used to report and interpret results. Do not report or interpret results in the code.

(20$^{\text{pts}}$) **1. Type conversions**
Type the following code in your R terminal.

```
v <- factor(c("2", "3", "5", "7", "11"))
```

(a) (5 pts) Describe the variable type/class of `v`.
(b) (5 pts) Convert `v` to character with `as.character()`. Explain what just happened.
(c) (5 pts) Convert `v` to numeric with `as.numeric()`. Explain what just happened.
(d) (5 pts) How would you convert the values of `v` to integers? Do it.

(70$^{\text{pts}}$) **2. From raw to technically correct data**
In this exercise we'll use `readLines()` to read in an irregular textfile: `http://statacumen.com/teach/ADA2/ADA2_HW_18_example.txt`. The file looks like this.

```
// Survey data. Created : 22 April 2015
// Field 1: Gender
// Field 2: Age (in years)
// Field 3: Weight (in kg)
M;28;81.3
male;45;
Female; 17 ;57,2
fem.;64;62.8
Ma.;16;55.3
m;;50,1
w;20.4;55
Fm;;
;55;55
```

First, we will read the data, work with the commented lines, then put the data lines into a matrix with column labels. Then, we will coerce the columns of the data to a structured data set.

Imagine that these three fields and four data rows are the first of potentially dozens of fields and 10,000 rows of data, so your strategy should be general and handle unseen but reasonable data values (for the columns in this dataset).

This is the type of coding that demands detailed comments for why you're writing each line of code and what your strategy is. Therefore, before each line of code, indicate the why and how of your code.

(a) (5 pts) Read the complete file using readLines.
(b) (10 pts) Separate the vector of lines into a vector containing comments and a vector containing the data. Hint: use `grepl()`.
(c) (5 pts) Extract the date from the first comment line and print to the console with the text "This data was created YYYY-MM-DD.", where the date is entirely numeric.
(d) (15 pts) Read the data into a matrix as follows.
    1. Split the character vectors in the vector containing data lines by semicolon (`;`) using `strsplit()`.
    2. Find the maximum number of fields retrieved by `split()`. Append rows that are shorter with `NA`'s.
    3. Use `unlist()` and `matrix()` to transform the data to row-column format.
(e) (5 pts) From comment lines 2-4, extract the names of the fields. Set these as `colnames` for the `matrix` you just created.
(f) (5 pts) Coerce the `matrix` to a `data.frame`, making sure all columns are `character` columns.

(g) (10 pts) Use a string distance technique to transform the `Gender` column into a `factor` variable with labels `man` and `woman`.

(h) (5 pts) Coerce the `Age` column to `integer`.

(i) (5 pts) Coerce the `weight` column to `numeric`. Hint: use `gsub()` to replace comma's with a period.

(j) (5 pts) Show off your beautiful dataset!