STAT 4410/8416 Homework 4

lastName firstName

Due on Nov 8, 2018

1. **Exploring XML data;** In this problem we will read the xml data. For this we will obtain a xml data called olive oils from the link http://www.ggobi.org/book/data/olive.xml. Please follow the directions in each step and provide your codes and output.

a. Parse the xml data from the above link and store in a object called `olive`. Obtain the root of the xml file and display its name.

b. Examine the actual file by going to the link above and identify the path of categorical variables in the xml tree. Use that path to obtain the categorical variable names. Please keep the names, not nick names and store them in `cvNames`. Display `cvNames`.

c. Now examine the file by going to the link and identify the path of real variables in the xml tree. Use that path to obtain the real variable names. Please keep the names, not nick names and store them in `rvNames`. Display `rvNames`.

d. Notice the path for the data in xml file. Use that path to obtain the data and store the data in a data frame called `oliveDat`. Change the column names as you have obtained the column names. Display some data.

e. Generate a plot of your choice to display any feature of `oliveDat` data. Notice that the column names are different fatty acids. The values are % of fatty acids found in the Italian olive oils coming from different regions and areas.

f. Explain what these two lines of codes are doing.

```
r <- xmlRoot(olive)
xmlSApply(r[[1]][[2]], xmlGetAttr, "name")
```

2. **Working with date-time data;** The object `myDate` contains the date and time when this question was provided to you. Based on this object answer the following questions.

```
myDate <- "2018-10-30 19:50:21"
```

a. Convert `myDate` into a date-time object with Chicago time zone. Display the result.
b. Write your codes so that it displays the week day of `myDate`.
c. What weekday is it after exactly 100 years from `myDate`? Show your codes and the answer.
d. Add one month with myDate and display the resulting date time. Explain why the time zone has changed even though you did not ask for time zone change.
e. Suppose this homework is due on November 8, 2018 by 11.59PM. Compute and display how many minutes you got to complete this homework?
f. Produce code that generates a numerical sequence representing every year between 1715 and 2018 inclusive, stored in `myYears`. Determine which of these years are leap years and store them in `leapYears`.
g. Report the number of 4-year and 8-year gaps in your `leapYears` data. Display all leap years after which an 8-year gap occurred until the next leap year. Of these years, how many days occurred between December 31st of the earliest and January 1st of the latest?

3. **Creating HTML Page;** In this problem we would like to create a basic HTML page. Please follow each of the steps below and finally submit your HTML file on the blackboard. Please note that you don't need to answer these questions here in the .Rmd file.
    a. Open a notepad or any plain text editor. Write down some basic HTML codes as shown in online (year 2014) Lecture 15, slide 6 and modify according to the following questions. Save the file as hw4.html and upload on the blackboard as a separate file.

b. Write "What is data science?" in the first header tag, `<h1></h1>`
c. Hw1 solution contains the answer of what is data science. The answer has three paragraphs. Write the three paragraphs of text about data science in three different paragraph tags `<p></p>`. You can copy the text from hw1 solution.
d. Write "What we learnt from hw1" in second heading under tag `<h2></h2>`
e. Copy all the points we learnt in hw1 solution. List all the points under ordered list tag `<ol></ol>`. Notice that each item of the list should be inside list item tag `<li></li>`.
f. Now we want to make the text beautiful. For this we would write some CSS codes in between `<head></head>` tag under `<style></style>`. For this please refer to online (year 2014) lecture 15 slide 8. First change the fonts of the body tag to Helvetica Neue.
g. For the paragraph that contains the definition of data science, give an attribute `id='dfn'` and in CSS change the color of 'dfn' to white, background-color to olive and font to be bold.
h. For other paragraphs, give an attribute `class='cls'` and in CSS change the color of 'cls' to green.
i. Write CSS so that color of h1,h2 becomes orange.
j. Write javaScripts codes so that onClick on `h1` header, it shows a message 'Its about data science'.

4. **Boston hubway data;** This question will explore Boston hubway data. Please carefully answer each question below including your codes and results.
   a. Obtain the compressed data, bicycle-rents.csv.zip, from blackboard and display few data rows.
   b. For each day, count the number of bikes rented for that date and show the data in a time series plot.
   c. Based on the rent date column, create two new columns weekDay and hourDay which represent week day name and hour of the day respectively. Store the data in myDat and display few records of the data. Hint: For weekday use function wday().
   d. Summarize myDat by weekDay based on the number of rents for each weekDay and store the data in weekDat. Display some data.
   e. Create a suitable plot of the data you stored in weekDay so that it displays number of bike rents for each week day.
   f. Now we want to investigate what happens in each day. Summarize myDat again but this time by weekDay and hourDay and obtain the number of rents. Store the data in hourDat and Display some data.
   g. The dataframe hourDat is now ready for plotting. Generate line plots showing number of bike rents vs hour of the day and colored by weekDay.
5. **Bonus for undergraduate (3 points) mandatory for graduate students:** The following link contains the complete texts of Romeo and Juliet written by Shakespeare. Read the complete text and generate a plot similar to Romeo and Juliet case study in online(year 2014) lecture 13 (last plot).

http://shakespeare.mit.edu/romeo_juliet/full.html

6. **Bonus (2 points) question for all** : In the United States, a Consumer Expenditure Survey (CE) is conducted each year to collect data on expenditures, income, and demographics. These data are available as public-use microdata (PUMD) files in the following link. Download the data for the year 2016 and explore. Provide some plots and numerical summary that creates some interest about this data.

https://www.bls.gov/cex/pumd.htm