# DA5020 – Assignment 5

In this assignment we will practice working with data in XML format and also the process of tidying data. You will work with two datasets: 1) the contact information for US Senators and 2) the Ratio Of Female To Male Youth Unemployment Rate.

You may want to complete the Data Camps String Manipulation with stringr in R &Intermediate Regular Expressions in R before attempting this assignment. Completing these data camps will count toward bonus points.

## Bonus — (+5 bonus points, *All instructions must be followed)*

1. Complete the data camp " String Manipulation with stringr in R" & a separate data camp "Intermediate Regular Expressionsin R ". A link to Data Camp will be shared within the course, so that you can access the resource for free using your Northeastern email. Please submit the statement of accomplishment from both data camps to earn credit (must also have your name on it). (No partial credit)

## Question 1 — (10 points)

Load the XML data directly from the URL below into a data frame (or tibble) and display the dimensions of the data.
URL: https://www.senate.gov/general/contact_information/senators_cfm.xml.
If you get errors while loading the XML file, ensure that you use the RCurl package.

**Use the XML data to answer questions 2 & 3 below.**

## Question 2 — (10 points)

Construct a regular expression (regex) to extract only the first, last names, and party (D,R,I) of each senator; the regex should remove their middle initial and/or suffix e.g. remove Jr. III, etc. Ensure that the updated names are reflected in the dataframe.

## Question 3 — (10 points)

Create a function called senatorsByState() which takes the two letter abbreviation for a US State as its input argument and displays the **first name, last name** and **party** of each senator for the selected state. For example, if the user enters 'MA' your function should display a message that is similar to the following: "The senators for MA are: Edward Markey, Democratic Party and Elizabeth Warren, Democratic Party"

**Answer the questions below using the attached dataset on the Ratio Of Female To Male Youth Unemployment Rate.**

## Question 4 — (10 points)

Download the attached csv file from Canvas and load it in your R environment. Perform steps to tidy the data and the prepared data should be divided across two tibbles named **country_name** and **indicator_data.** The **country_name** tibble should contain the country name and country code (ensure that you remove duplicates), and the **indicator_data** tibble should include the country_code, year, and value. Note: Tidy the data using pivot_longer(), pivot_wider() and separate(), where applicable.

## Question 5 — (10 points)

Select five countries from each of the following continents: North America, Asia and Middle East. Visualize their respective data from the last 20 years using a line chart; use facet_wrap to display the data for each continent in its own chart. Explain your observations about the ratio of female to male youth unemployment rate in the selected regions.

## Submission Details

- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.A5.FirstName.LastName.Rmd and your PDF/HTML DA5020.A5.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.
- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it.) Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.
- Not submitting a knitted PDF or HTML will result in reduction of 30 points.
- Not submitting the .Rmd file (or both) will result in a score of 0.

## Useful Resources

- R Markdown Notebooks
- XML file containing the contact information of US Senators Contact.
- World Bank Ratio Of Female To Male Youth Unemployment Rate

** Note: Do not merely print the output for each question but always explain the results from your code. Points will be deducted for any missing explanation.