# MSDS 6306: Doing Data Science

## Live session Unit 03 assignment

## Due: 1 hour before your 4th live session

### Submission

**ALL MATERIAL MUST BE KNITTED INTO A SINGLE, LEGIBLE, AND DOCUMENTED HTML DOCUMENT.** Formatting can be basic, but it should be easily human-readable. Unless otherwise stated, please enable {r, echo=TRUE} so your code is visible.

### Questions

1. **GitHub Cloning (20 points):** Using Git, clone the following GitHub repository to your local machine: https://github.com/caesar0301/awesome-public-datasets. In RMarkdown, please show the code (commented out, as it's not R syntax) that you used to create a new directory, navigate to the appropriate directory, and clone the repository to it. One Git command per line, please.

2. **Data Summary (20 points):** From this aforementioned cloned repo, please extract titanic.csv.zip. To be clear, this does not have to be done in Git or command line.

    a. In R, please read in titanic.csv via either read.table() or read.csv(), assigning it to df. This dataset follows the passengers aboard the Titanic, including their fees paid, rooms rented, and survivorship status.

    b. Output the respective count of females and males aboard the Titanic. Plot the frequency of females and males. Be sure to give an accurate title and label the axes.

    c. Please use one *apply* function (to review: swirl() modules 11, 12) to output the means of Age, Fare, and Survival. Make sure the output is a real number for all three means.

3. **Function Building (30 points)**: You research sleep and just got your first data set. Later, you'll have another dataset with the same column names, so you want to create a helper function that you can analyze this dataset and the next. Load sleep_data_01.csv (found at http://talklab.psy.gla.ac.uk/L1_labs/lab_1/homework/index.html). Questions 3A through 3D should be answered in function(x){}. 3E can be outside of the function.

    a. Create objects for the median Age, the minimum and maximum Duration of sleep, and the mean **and** standard deviation of the Rosenberg Self Esteem scale (RSES). You may need to specify a few options like in Problem 2 and live session.

    b. Create a data.frame object called report: it should consist of the median age, the RSES mean **and** standard deviation respectively divided by five (since there are five questions and these scores are summed), and the range of Duration (the statistical definition of range; it should be a single number.)

**c.** Change the column names of this data.frame to MedianAge, SelfEsteem, SE_SD, and DurationRange.

**d.** Round the report to at *most* 2 digits: leave this as the closing line to the function.

**e.** Finally, run the function on your sleep data to show the output.

4. **FiveThirtyEight Data (30 points):** Navigate on GitHub to https://github.com/rudeboybert/fivethirtyeight and **read README.md**. It will include everything you need.

   **a.** Install the fivethirtyeight package.

   **b.** In the *listing of Data sets in package 'fivethirtyeight,'* assign the 22nd data set to an object 'df.'

   **c.** Use a *more detailed list of the data sets* to write out the URL in a comment to the related news story.

   **d.** Using R command(s), give the dimensions and column names of this data frame.

5. **Data Summary (30 points):** Use your newly assigned data frame from question 4 for this question.

   **a.** Write an R command that gives you the column names of the data frame. Right after that, write one that counts the number of columns **but not** rows. **Hint:** The number should match one of your numbers in Question 1d for dimensions.

   **b.** Generate a count of each unique major_category in the data frame. I recommend using libraries to help. To be clear, this should look like a matrix or data frame containing the major_category and the frequency it occurs in the dataset. Assign it to major_count.

   **c.** To make things easier to read, put par(las=2) before your plot to make the text perpendicular to the axis. Make a barplot of major_count. Make sure to label the title with something informative (check the vignette if you need), label the x and y axis, and make it any color other than grey. Assign the major_category labels to their respective bar. Flip the barplot horizontally so that bars extend to the right, not upward. All of these options can be done in a single pass of barplot(). **Note:** It's okay if it's wider than the preview pane.

   **d.** Write the fivethirtyeight data to a csv file. Make sure that it does not have row labels.

6. **Codebook (30 points)**:

   **a.** Start a new repository on GitHub for your SMU MSDS homework. On your local device, make sure there is a directory for Homework at the minimum; you are welcome to add whatever you would like to this repo in addition to your requirements here.

   **b.** Create a README.md file which explains the purpose of the repository, the topics

included, the sources for the material you post, and contact information in case of questions. Remember, the one in the root directory should be general. You are welcome to make short READMEs for each assignment individually in other folders.

**c.** In one (or more) of the nested directories, post your RMarkdown script, HTML file, and data from 'fivethirtyeight.' Make sure that in your README or elsewhere that you credit fivethirtyeight in some way.

**d.** In your RMarkdown script, please provide the link to this GitHub so the grader can see it.

**Reminder**

To complete this assignment, please submit **one** RMarkdown and matching HTML file that includes questions 1-6 at least one hour before your 4th live session. Please submit all files at the same time; only one submission is granted.

Good luck!