**MSDS 6306: Doing Data Science – File Management**

Live session Unit 03 assignment

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

**Submission**

**ALL (non-swirl) 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**

• **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.

**• 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.

• 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.

• 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.

• 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.

**• 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.

• 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.

• 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.)

• Change the column names of this data.frame to MedianAge, SelfEsteem, SE\_SD, and DurationRange.

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

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

**• Swirl (30 points)**: Complete Modules 12 to 14 in the R Programming course of Swirl. *Copy your code/output to a separate .txt file. It does not need to be included in your RMarkdown file. The grader has requested at minimum to show the 90%-100% progress bar for each Module and what output you had for it.*

• Complete “12: Looking at Data”

• Complete “13: Simulation”

• Complete “14: Dates and Times”

**Reminder**

To complete this assignment, please submit **one** RMarkdown and matching HTML file that includes questions 1-3, and a .txt file containing solely your swirl output (Question 4) at least one hour before your live session. Please submit all files at the same time; only one submission is granted.

Good luck!