



## **L1: R Refresher Lab (10 points)**

Instructor: Nolan Bett

TA: Abdulateef Gafar & Wenhuan Xu

**DUE DATE:** Friday, January 20<sup>th</sup> @ 9pm

### **OVERVIEW OF LAB:**

This lab is designed as a refresher (or perhaps an introduction) to R, to re-acquaint you with the software and how to use it for data analysis. R is an increasingly common (and ever evolving) open-source tool used in academia and industry that can perform very sophisticated tasks in data and statistical analysis. To code in R we will rely on RStudio, which is an IDE (Integrated Development Environment).

The lab exercise contains simple tasks focused on data organization, data manipulation, graphics, and descriptive statistics. In this lab we'll use simple datasets that have been nicely curated. Not all data you encounter will be this clean or nice.

We have also put together a comprehensive [CONS 452 R Wiki](#) website that has lots of helpful information for how to use R. It's a good place to start if you need help with something. We will also post some R guides on the Canvas website in the 'How-to-Guides' section.

In this lab you'll be working from an RScript that we have provided. Make sure you input your code where it is required (you will need to submit the completed code), and answer all the questions (see below) in a separate Word document.

### **LEARNING OBJECTIVES:**

By the end of this exercise, you should be familiar with:

- RStudio as an interface: know how define your working directory, start a new R script
- Importing and Exporting data:
  - Import data into RStudio
  - Save/export data as tables(csv), images, or html
- Data manipulation:
  - Add or remove columns

- Filter data based on specific criteria
  - Perform simple algebra on data values
- Describing your data:
  - Mean/mode/median
  - Standard Deviation
  - Contingency tables
- Data visualization:
  - Basic plotting functions (scatter plots, box plots, frequency plots)
  - Get familiar with ggplot (a useful plotting function)

**DELIVERABLES:**

Please upload two documents to Canvas:

1. A word file or pdf with your answers to the questions below, including the graphs created in this lab (please include good captions for each graph)
2. The R-script that we gave you, with the blanks filled in (by you)

**BEFORE YOU BEGIN:**

1. Create a new folder on your computer's Desktop and name it `CONS_452_YOUR INITIALS`. Create a sub folder and call it `L1_Intro_to_R`
2. Download the `titanic.csv` and `rainforest.csv` files and save them to the `L1_Intro_to_R` folder
3. Open the R script provided in RStudio

**QUESTIONS:**

Overall organization/formatting of your word/pdf file. (0.5 points)

Q1: What is the titanic dataset about? Name the variables and describe what information they provide. (1 point)

Q2: How many passenger classes did the Titanic have? (0.5 points)

Q3: In a life-threatening situation, conventionally the "Women and children first" policy is followed when deciding who to save first. Can you use the table commands you just learned to check whether that policy was followed, as the Titanic sunk (just make sure the policy was followed with regards to Women and Men)? Describe your answer quoting necessary proportion numbers. (2 points)

Q4: Using the commands you just tried, recreate plots 1,2, and 4. Make sure to label the X and Y axes. A title above the plot is not needed. Save them with appropriate file names in your Lab 1 folder. Insert the plot into your final submission, and below each plot write an appropriate caption (you can write the caption directly into your Word Doc). See the note below about writing captions. (1 point per plot)

Q 5: Create (or recreate) plots 5, 6 and 7. Add appropriate labels and make them visually appealing (proper font size, etc), then save them as separate files. Paste the saved plots into your submission document. Give them appropriate captions. (1 point per plot).

**Note On Captions:**

A good caption orients the reader to the figure. It should be stand-alone--that is, the reader should be able to understand the figure without having to refer to the rest of the paper. The first sentence usually explains the relationship that it shows, often as a statement that summarizes the main trend shown in the figure. Then, there is often a sentence about the underlying data if this is needed to fully understand the figure (perhaps how the data were collected or manipulated). If there were any statistics conducted, then there is usually also a sentence highlighting the statistical results. For this assignment, see if you can start by making a simple caption that describes what the figure is showing and/or summarizes the main relationship between the X and Y variables.