**Introduction to R/Posit (Formerly known as RStudio), I**

1. **Pre-work:** 
   * You will use **Google Cloud Platform (GCP)** for in-class work. At least **1 week before the course**, R-IT will add you to GCP and will email you. Follow their instructions to connect.
   * **Optional (but recommended)**: Install R/Posit locally—especially useful after the course ends. You can request both R and RStudio from IT [**here**](https://jacksonlaboratory.sharepoint.com/sites/IT/Lists/Software%20Catalog%202/AllItems.aspx?e=HvrPC6&CID=a8c5ecbc%2D26d5%2D407f%2D8180%2Dae2feaceda5f)**.**
2. **Why R:**
   * Like Python, it is a “Table Stakes” language for Data Science
   * Has access to many important libraries for genomics/bioinformatics and visualizations
   * <https://www.oreilly.com/radar/2022-cloud-salary-survey/>
3. **Course Description:**

This workshop is the first of two workshops that constitute a broad survey of the R programming language, a major language central in the fields of bioinformatics, genomics, statistics, and Data Science. No prior programming experience is required. We will use RStudio (now called “Posit”) and R installed on Google Cloud Platform (GCP). If you would like to continue to use R and RStudio on your own laptop afer the course ends, then **you should also have both of these applications installed prior to the beginning of the course on your laptop since the cloud instance will be shut down when the course ends. You can request R and RStudio** [**here**](https://jacksonlaboratory.sharepoint.com/sites/IT/Lists/Software%20Catalog%202/AllItems.aspx?e=HvrPC6&CID=a8c5ecbc%2D26d5%2D407f%2D8180%2Dae2feaceda5f) **and have IT install them for you**. Please note that downloading RStudio will not simultaneously install R - you need to install ***both, not just one***. After completion of this initial class, there will be a second half offered later in the year (R II) which will cover packages important in Data Science (e.g. tidyverse) and basic statistical analysis capabilities of R.

1. **Course Goals:**

This course will provide you with the tools to analysis data. By the end of the week, you will be able to:

* Navigate the Posit (RStudio) interface and use the RMarkdown Language to document analysis
* Understand the syntax and structure of R
* Understand and apply core R data structures: vectors, lists, data frames, and matrices.
* Create basic scripts in R focused on importing, transforming, and visualizing data
* Use built-in tools to visualize data in a reproducible way
* Use ggplot2 to visualize data

1. **Locations of the course:**

* **See Module 0 for zoom link, passcode and for classrooms for those attending in Bar Harbor.**

1. **Instructor:** Danielle Presgraves, Ph.D. **email:** [Danielle.presgraves@jax.org](mailto:Danielle.presgraves@jax.org)

|  |  |  |
| --- | --- | --- |
| Module  Est. time | Topic Covered | Enrichment Suggestions |
| **0**  **<1 hour** | **Activities:**   1. Respond to IT when they add you to the GCP 2. (optional) Download R/RStudio (now “Posit”) and ensure that everything is working prior to class.   **Instructional and Assessment Activities:**   1. Please fill out the brief survey | 1. General R:  * <https://r4ds.hadley.nz/> * <https://intro2r.com/>  1. Statistics Summary (flowchart):   https://www.statsflowchart.co.uk |
| ***RStudio (Posit) Fundamentals*** | | |
| **1A**  **2 hours** | **Title:** Introduction to R and RStudio and markdown  **Learning Objectives:**  You will demonstrate a minimum level of competency in the RStudio environment. Competency includes:   * You will create a new .rmd file * You will correctly incorporate the R Markdown language * You will appropriately use 'R chunks' to manipulate data using R * You will successfully 'Knit' (equivalent to 'compiling') your final .rmd document into a word document 🡪 save as a html or word * You will examine the basic features of dataframes * You will manipulate a built-in data set as well as upload an external .csv file into the RStudio memory   **Instructional and Assessment Activities:**   1. We will work through Module\_1A 2. We will import a dataframe and peek at it. 3. We will create, save, and knit our own file for our first Assignment. | **R Markdown Intro Cheat sheet:**  <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>  How to connect to the HPC for R document (see Module 0)  **Outside class time Activity for credit:**   * Day 1 Assignment (see below) * Topics in Assignment: * Write .rmd and knit * Analyze and reshape dataset using core |
| ***Manipulating and Refining Datasets*** | | |
| **1B**  **1.5 hours** | **Title:** Data Types & Structures in R  **Learning Objectives:**   1. Data Structures in R    1. Vectors, factors, matrices, lists, dataframes 2. You will download, access, and manipulatedata    1. Subset, tapply   **Instructional and Assessment Activities:**   1. We will work through Module\_1B 2. We will compare and contrast vectors, dataframes and different types of data.   **Assignment:**   * Module 1 Assignment (See Canvas) |  |
| ***R Programming Fundamentals*** | | |
| **2AB**  Instruction:  **2.5 hours** | **Title:** Programming in R  **Learning Objectives:**   1. Identify built-in functions in R 2. Understand the syntax of defining your own function, including anonymous functions. 3. You will use the basic components of programming: loops, defining their own function, demonstrate competency with a subset of higher order functions such as 'filter', and ‘map’ (for iterative tasks) 4. (optional): swirl() package   **Assignment:**   * Module 2 Assignment (See Canvas) | \*apply() family of functions:  <https://ademos.people.uic.edu/Chapter4.html>  **Outside class time Activity for credit:**   * Day 2 Assignment (see below) * Topics in Assignment: * Write small functions that clean and summarize data |
| ***Returning to Manipulating and Refining Datasets*** | | |
| **3AB**  **2 hours** | **Title:** Data Wrangling: part 1  **Learning Objectives:**   1. Slicing 2. Logical/ Boolean operators 3. Built in functions: which, subset, \*apply family   **Assignment:**   * Module 3 Assignment (See Canvas) | **Outside class time Activity for credit:**   * Day 3 Assignment (see below) * Topics in Assignment: * Clean and format a messy dataset for analysis |
| ***Using R for Descriptive Statistics & Visualization*** | | |
| **4A**  **1.5 hours** | **Title:** Describing Uncertainty: Confidence intervals  **Learning Objectives:**   1. You will visualize data using the appropriate display format: scatterplot, histogram, stem and leaf plots 2. You will investigate the benefits of boxplots for visualizing uncertainty in data sets 3. (Optional) You will simulate a bootstrap to calculate quantiles of data sets | **Outside class time Activity for credit:**   * Day 4 Assignment (see below) * Topics in Assignment: * Explore a new dataset and visualize using standard tools |
| **4B**  **1.5 hours** | **Title:** Graphics and Basic Statistics mosaic plots, scatterplots  **Learning Objectives:**   1. You will visualize data using the appropriate display format: scatterplot, histogram, stem and leaf plots   **Assignment:**   * Module 4 Assignment (See Canvas) | **Additional resources:**   1. <https://r4ds.hadley.nz/> 2. [**https://www.youtube.com/watch?v=sxQaBpKfDRk**](https://www.youtube.com/watch?v=sxQaBpKfDRk) 3. [**https://www.data-to-viz.com/#boxplot**](https://www.data-to-viz.com/#boxplot) 4. [**http://www.cookbook-r.com/Graphs/**](http://www.cookbook-r.com/Graphs/) 5. [**https://tidydatatutor.com**](https://tidydatatutor.com) |
| **5A**  **2 hours** | **Title:** The 'grammar of visualization': ggplot2  **Learning Objectives:**   1. You will discover how to import libraries! 2. You will examine the basic model of ggplot2 3. You will compare and contrast the data visualizations in module 2 to the more attractive results of the ggplot2 package   **Assignment:**   * Module 5 Assignment (See Canvas) | **Outside class time Activity for credit:**   * Day 5 Assignment (see below) * Topics in Assignment: * Explore a new dataset and visualize using ggplot2 to compare to standard R tools |
|  |  |  |

**Day 1 Assignment:**

Hand in a .rmd file saved as “***FirstInitialLastname\_Day1.rmd***” that has ***BOTH markdown explanations and coding chunks (with hashed out comments).***

1. Use an **R Markdown chunk (not a coding chunk)** to introduce yourself. This doesn't have to be long but should contain at least one list and one header, so you become familiar with how to use Markdown (and how to use the internet/cheatsheets to help you!). As a prompt, you might answer a subset of the following (or anything other similar question): Why do you want to learn R? What is your background? What are your favorite foods or Hobbies? Do you have pets? If so, why are they THE best? Gimme me an enumerated list of the five things that (a) make you happiest, OR (b) top five people you would invite to a dinner party, OR (c) top five books or films that you would recommend.
2. Let’s say that in our experimental analyses, we are working with three different sets of cells: normal, cells knocked out for geneA (this means they are missing geneA), and cells overexpressing geneA. We have three replicates for each of the three cell types (normal, none, over-expressed).
3. Create a vector named samplegroup with nine elements: 3 control (“CTL”) values, 3 knock-out (“KO”) values, and 3 over-expressing (“OE”) values
4. Turn samplegroup into a factor data structure.

1. Turn the following table into a matrix (You can just use the Weight and Height columns for your matrix):

|  |  |  |  |
| --- | --- | --- | --- |
| Species | Sex | Weight (lbs) | Height (cm) |
| Gorilla gorilla | M | 400 | 180 |
| Gorilla gorilla | F | 175 | 140 |
| Pan troglodytes | M | 120 | 150 |
| Pan troglodytes | F | 85 | 150 |
| Homo sapiens | M | 180 | 178 |
| Homo sapiens | F | 140 | 165 |
| Orangutan | M | 160 | 137 |
| Orangutan | F | 82 | 115 |
| Pan paniscus | M | 86 | 78 |
| Pan paniscus | F | 68 | 73 |

1. Using the Bumpus\_data.csv:
2. Read in the Bumpus data set.
3. Pick one of the remaining columns and conduct analysis on it using tapply (which, recall, is for numeric continuous variables). You can use built in functions like sd() or mean() or some other built-in function that sparks joy (you can google lists of built-in base R functions, but you can also go here: https://www.javatpoint.com/r-built-in-functions). Use the R Markdown chunks to explain your analysis (Why did you pick this function? What is the data type of the column? Was the output what you expected?)

**Day 2 Assignment:**

Hand in a .rmd file saved as “FirstInitialLastname\_Day2.rmd” that has markdown explanations and coding chunks (with hashed out comments). It is not cheating to use google for help!

1. Write a function that takes one number and raises it to the power of a second number.
2. Write a function called multiply\_it, which takes three inputs: a numeric value x, a numeric value y, and a numeric value z. The function will return the product of these three numeric values, which is x \* y\* z. For example, multiply\_it(4, 6,2) will return output 48. At least one of the three passed arguments should have a default value.
3. Call the following with an anonymous function that takes one number and raises it to the power of itself (this solution should be similar to Q1):

**evaluate <- function(func, dat){**

**return(func(dat))**

**}**

1. Write out**two different** ways of creating a vector that only contains the even numbers between 0 and 20 (You could use a higher order function, a built-in function, or you could define your own function).
2. Use a higher order function to create a vector that multiplies each element by itself. Print out the resulting vector to the screen

**Day 3 Assignment:**

Hand in a file name "firstinitialLastname\_Day3.rmd".

1. Using the Great Apes Dataframe, do the following:
   1. With slicing: Reverse the rows so that Pan Paniscus are the first two rows
   2. With slicing: Print out all the columns except for the height column
   3. With slicing: Only include males that are smaller than 120 lbs or shorter than 140 cm
   4. With Slicing: Add a column to your dataframe of whatever characteristic you choose (age of first reproduction, number of chromosomes etc.).
   5. Use the subset function to reproduce c.
   6. Now that you have added a column to the Great Apes dataframe, save it (under a different name) to your RAssignments folder.
2. Using the Bumpus data set: use tapply to investigate a summary statistic (i.e. mean, median, sd, var) over the column of your choosing but ensure that your information is split by **sex**and **survival**.

**Day 4 Assignment:**

Hand in a .rmd file saved as “FirstInitialLastname\_Day4.rmd” that has markdown explanations and coding chunks (with hashed out comments).\

**USE BASE R to do the following (you will redo the same question using ggplot2 for Module 5 Assignment).**

With the Bumpus dataset choose a numeric trait (one of the columns such as head, humer, wgt etc.):

1. Using two different types of graphs, visually inspect to determine if your chosen variable is normally distributed.
2. Graph a boxplot for your chosen variable that is divided by sex and survival.
3. Draw a mosaic plot of the survival of the female versus male birds.

**Day 5 Assignment:**

Hand in a .rmd file saved as “FirstInitialLastname\_Day5.rmd” that has markdown explanations and coding chunks (with hashed out comments).

**Redo your visualizations from Module 4, but use ggplot instead of Base R. We did not cover all types of graphs, but armed with google (and a cheatsheet), this should be reasonably straightforward to translate within the ggplot layers.**

With the Bumpus dataset choose a numeric trait (one of the columns such as head, humer, wgt etc.):

1. Using two different types of graphs, visually inspect to determine if your chosen variable is normally distributed.
2. Graph a boxplot for your chosen variable that is divided by sex and survival.
3. Draw a mosaic plot (Note: use a stacked bar plot scaled to 1, same as the diamond example in Module 5A, to mimic a mosaic plot) of the survival of the female versus male birds.