Introduction to RStudio

This first data analysis assignment is intended to help you familiarize yourself with the RStudio interface and get comfortable running small chunks of R code. If you have not done so already, please work through the Introduction to RStudio & Tutorial in the Getting Started & Course Resources module on Canvas.

For this assignment, you will need the loan50.csv dataset which can be found on the Data Analysis 1 assignment page and in the Introduction to RStudio & Tutorial.

You can find a description of the variables recorded in the loan50.csv dataset on the [OpenIntro Statistics loan50 info page](https://www.openintro.org/data/index.php?data=loan50).

The Introduction to RStudio & Tutorial walks you through using some of the basic, built-in functions in R. Read through and run each line of code to ensure you understand what the functions are doing and what types of output each produces.

# Part 1: Present Day Birth Records in the United States

In Part 1 of the Introduction to RStudio & Tutorial, you worked through visualizing and summarizing data from the Arbuthnot’s baptism data. This assignment involves repeating these steps, but for present day birth records in the United States The data are stored in a data frame called present.

Load the present dataset using the data() function, just as we did in the tutorial. Note that if you do not have the openintro package loaded, you will not be able to access this dataset. The tutorial discusses how to open packages using the library() function. Once you have the present dataset loaded, answer the following questions.

1. (1 point) What years are included in this dataset?
2. (1 point) What are the dimensions of the data frame? Dimensions refers to the number of rows and variables.
3. (1 point) What are the variable (column) names?
4. (1 point) Make a plot that displays the proportion of boys born over time. What do you see? Does Arbuthnot’s observation about boys being born in greater proportion than girls hold up in the U.S.? Include the plot in your response. *To copy or save a graph from RStudio, click the Export button just above the preview of the graph. From there you can choose to Save Image or Copy to Clipboard.*
5. (1 point) In what year did we see the most total number of births in the U.S.? *Hint:* First calculate the totals and save it as a new variable. Then, sort your dataset in descending order based on the total column. You can do this interactively in the data viewer by clicking on the arrows next to the variable names. To include the sorted result in your report you will need to use two new functions. First we use arrange() to sorting the variable. Then we can arrange the data in a descending order with another function, desc(), for descending order. The sample code is provided below.  
     
   present %>%  
   arrange(desc(total))

# Part 2: Loan Data from Lending Club

Part 2 of the Introduction to RStudio & Tutorial uses two variables from the Lending Club data: loan\_amount and homeownership. For the assignment you’ll submit, you will practice using two **different variables**. Please make sure the assignment you submit uses the correct variables (specified in the questions below).

## Exploring a Single Quantitative Variable

For this portion of the assignment, you’ll practice using R to explore the **annual\_income** variable in the loan50.csv data set.

a. (1 point) Construct a histogram of the annual income data. Include informative labels and a title. Include your histogram below.

b. (1 point) Construct a boxplot of the annual income data. Include informative labels and a title. Include your boxplot below.

c. (1 point) Using the histogram you constructed in part a and the boxplot from part b, describe the shape of the distribution of the annual income variable and comment on the presence of any outliers.

d. (1.5 points) Calculate the mean of the annual income data.

e. (1.5 points) Calculate the median of the annual income data.

f. (1 point) Which measure of center (mean or median) is more appropriate for these data? Why? Consider the shape of the distribution discussed in part c.

g. (1.5 points) Calculate the standard deviation of the annual income data.

h. (1.5 points) Calculate the interquartile range of the annual income data.

## Visualizing Two Variables

Let’s continue to explore the annual income data, but now consider how annual income data may vary between loan status (current or fully paid).

i. (1 point) Construct a side-by-side boxplot for annual income broken up by loan status. Include informative labels and a title.

j. (2 points) How do the distributions of annual income compare for loan status? Comment on the shape, center, spread, and presence of outliers for the two groups.

## Exploring a Single Categorical Variable

Finally, we’ll focus our attention only on the loan status variable.

k. (2 points) Construct a table of counts for the loan status variable. Report the number of observations in each category below.

l. (2 points) Construct a table of proportions for the loan status variable. Report the proportions for each category below.

m. (1 point) Construct a barplot that displays the distribution of loan status types. Include informative labels and a title. Include your barplot below.

**Gradescope Page Matching (2 points)**

When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page “Submitting Assignments in Gradescope”. Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item “Gradescope Page Matching”.