Introduction to data/Introduction to R

STAT 140

Marie Ozanne

September 4-6, 2019

## Class Activity - Introduction to data

For this activity, you will collect information from each of your classmates and store those data in an Excel spreadsheet. Find out the name, age, class year, height, and two other pieces of information about *each* of your classmates. Make sure you collect the same information on each person and that you include the same information about yourself. Use your best judgement on how to organize what you collect - you will not be assessed on whether you are “right” or “wrong”, but rather we will use this exercise as an opportunity to explore how to store data in this class.

## Class Activity - Introduction to R

### Getting Started

1. Download R: <https://cran.r-project.org/>
2. Download R Studio (Desktop, Open Source License): <https://www.rstudio.com/products/rstudio/download/>
3. Download this .Rmd file from the class website: <https://marievozanne.github.io/STAT_140_01_F19_notes.html> (click on the link in the “Code” column). The .Rmd extension denotes an R Markdown file - a file format that allows you to write reproducible reports with R. I will be using R Markdown for class notes and providing you with templates that you will use for your homework and class assignments.

### R Setup

When we use R Markdown, we can include a combination of text, equations, and R code in our document. In this activity, we will focus on learning a little bit about R and incorporating R code into the document. R code is included in an R Markdown file in a “chunk”.

1. Insert a chunk by pressing the Insert button above and selecting R:
2. Each chunk that you insert must have a unique name for your document to compile. One example is a chunk named “setup”:

Some of the functionality of R is built into your computer now that you have downloaded R and RStudio, but some of it relies on external files called “packages” that can be downloaded from CRAN R. If you need any packages to make plots, for example, you should load these files in your setup chunk. You can also load data sets in this chunk.

Notice that I named this chunk setup2, rather than setup. If you wish, you can rename it as setup and try to knit your document (press the Knit button above) - what happens?

There are some packages that we will need that do not come preloaded with R. You can install them by navigating to the Packages tab, hitting Install, and specifying the package that you want to install. This is only necessary for packages that are not built into R. I will tell you if you need to install a package, so don’t worry too much about this. Below, we will need to install ggplot2, a popular data visualization package. You only need to install ggplot2 once, so the next time you need to use this package (and you’re working on the same computer), you do not need to perform this step. After the package is installed, you load it using the library() function.

Whenever you write a chunk of code, you should make comments about what you are doing so it is easy for you and others to understand the code at a later date. If you wish to make a comment, precede it with (at least) one #. Notice the comments included in the chunk below.

## Install ggplot2 (only the first time you use this package)  
  
## Load package  
library(ggplot2)  
  
## Load data - Dr. Arbutnot's Baptism Records  
source("http://www.openintro.org/stat/data/arbuthnot.R")

### Exploratory analysis - Dr. Arbuthnot’s baptism records

* The command head() allows us to view the first few rows of a data set.
* When the data are stored in a data frame (a kind of spreadsheet or table), the command dim() returns the rows and columns, respectively, of the data set. It is important to check that these numbers match what you are expecting.
* It is important to know the names of the columns of the data set, since we will be organizing data such that the column names are the *variables* we care about (more on this later).
* The command str() returns a summary of the type of object (a data frame), the dimensions, and the variable names and types.
* We can look at relationships between variables graphically. We will use ggplot2 for this. The ggplot command creates plots using layers. First, within ggplot(), we specify the data set and within aes() we specify the x and y variables (aes refers to “aesthetics”). Then we add a layer to denote the kind of plot we want. Below you see an example of a line graph and a dot plot.
* You can access a particular variable in a data frame using the $.

## Look at the first few rows of a data set.  
## These data are stored in a data frame  
head(arbuthnot)

## year boys girls  
## 1 1629 5218 4683  
## 2 1630 4858 4457  
## 3 1631 4422 4102  
## 4 1632 4994 4590  
## 5 1633 5158 4839  
## 6 1634 5035 4820

## What is the size of the data set?   
## dim() returns the rows and columns, respectively, of the data set  
dim(arbuthnot)

## [1] 82 3

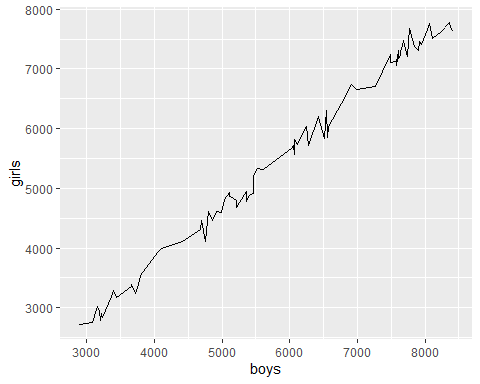
## What variables are in this data set?  
names(arbuthnot)

## [1] "year" "boys" "girls"

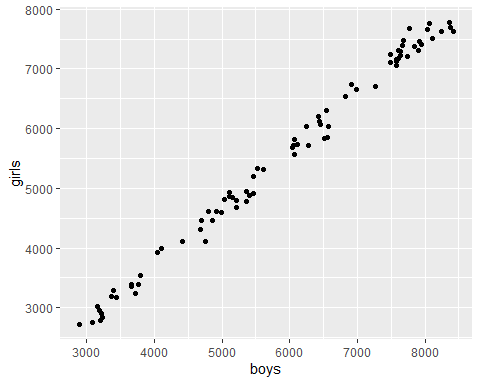
## What kinds of variables are in the data set?  
str(arbuthnot)

## 'data.frame': 82 obs. of 3 variables:  
## $ year : int 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 ...  
## $ boys : int 5218 4858 4422 4994 5158 5035 5106 4917 4703 5359 ...  
## $ girls: int 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 ...

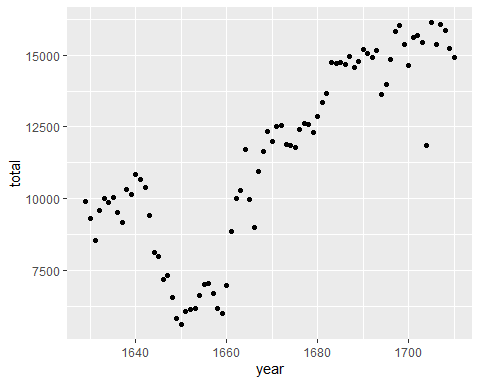
## Relationship between number of girls and boys baptized (line graph):  
ggplot(data=arbuthnot, aes(x=boys, y=girls)) + geom\_line()



## Relationship between number of girls and boys baptized (dot plot):  
ggplot(data=arbuthnot, aes(x=boys, y=girls)) + geom\_point()



## How many children were baptized each year?  
### The $ lets you access a particular column of a data set by name.  
### R computes all sums simulateously, so you don't have to add up   
### boys and girls for each year one by one.  
  
### You can create a new variable for a data set as follows:  
arbuthnot$total <- arbuthnot$girls+arbuthnot$boys  
  
ggplot(data=arbuthnot, aes(x=year, y=total)) + geom\_point()



### Class data

Now we are going to practice what we have learned using the data you collected on your classmates in the previous class. First, you need to read in your data from Excel. The step for this is different than what we used for the arbuthnot data because that data was stored on a website. Please make sure that your file is stored as a .csv file. This is the easiest file format to import in R, although you can import others. We will use the read.csv() function and store the data using a descriptive name, like “myclassdata”. Now we just need to specify the file path.

#### Setup

## You can use this line of code to read in your data - just remove the # in the next line.  
# myclassdata <- read.csv(file="", header=TRUE)

#### Main Questions

Answer the following questions about your class data by creating R chunks and implementing the appropriate code. You may work together if you get stuck, or you can ask me for help. Be sure to document what you are doing!

* Do the data look like you expect? What are the dimensions, variables, and variable types? Are there any missing values?
* What class years are represented? Hint, the unique() function will return the unique levels (here, class years) of a variable. Learn about this function by typing $unique() into the console.
* How many students are in each year? Hint, you should explore the table() function. You can learn about it by typing ?table into the console.
* What is the distribution of heights in the class? Hint, think about how you should store height so R can read it and look into the hist() function to examine this question graphically.
* Summarize any other information you collected that you think is interesting. Talk to me or to your classmates about how you would like to do this. Feel free to write up your ideas here but not implement them if you are having trouble figuring out how to write the R code.

#### Other Questions

* Why did we collect the names of all your classmates, aside from getting to know them?
* If you were going to share information you collected and wanted to keep it anonymous, what could you have reported in lieu of names?
* Did you collect any information that you think is especially hard to summarize (e.g., sentences)? Why is this information hard to summarize and can you think of any ways that you might do it?