Welcome to programming in R

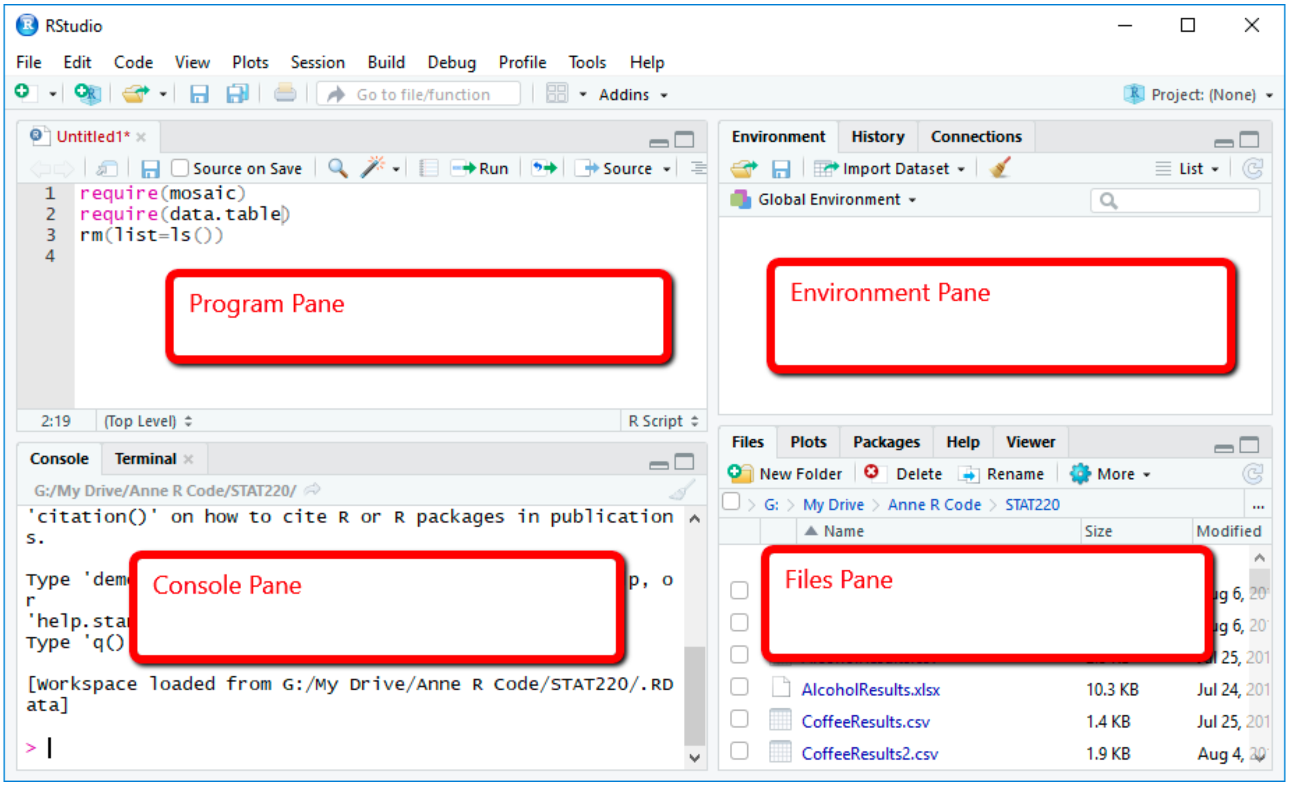
The goal of lab sessions is to develop your proficiency in the R language for statistical analysis, so that you are able to perform the statistical analyses and learn to visualize data using the techniques that we have learned during lecture.

R is a computational environment that is used for a wide range of statistical and mathematical analyses. For example, it can be used as a calculator; for creating stunning graphs; run basic and advanced statistics; numerically solve differential equations; and for more specialized analyses such as found in bioinformatics.

To install R on your personal laptops, go to cran.r-project.org (CRAN stands for the Comprehensive R Archive Network) and click on one of the MacOS X or Windows links in the "Download and Install R" box. Follow the links for your operating system (choose 'base files' for Windows users) and download the latest version ('R-3.6.2-win32.exe' for Windows users and 'R-3.6.2.dmg' for MacOSX users). The downloaded file has an installer in it, so double clicking the file and following the prompts is all you need to do to install R.

R Studio is an interactve development environment (IDE) for the R language. It allows programmers to execute their code in real time. While I personally do not use R Studio (I use a text editor linked to the R console), I find that it is a great tool for intuitive coding in R. To install R Studio on your personal machines, you will need to download the R Studio Installer from https://rstudio.com/products/rstudio/download/ Please follow instructions for installing R studio on the download page. Please note that you cannot install R studio without installing base R first.

To begin, please open R studio.



```
require(mosaic)
require(data.table)
rm(list=ls())
```

```
G:/My Drive/Anne R Code/STAT220/
'citation()' on how to cite R or R packages in publication
s.

Type 'dem                                        p, o
r
'help.sta
Type 'q()

[Workspace loaded from G:/My Drive/Anne R Code/STAT220/.RD
ata]

> |
```

Name	Size	Modified
AlcoholResults.xlsx	10.3 KB	Jul 24, 201
CoffeeResults.csv	1.4 KB	Jul 25, 201
CoffeeResults2.csv	1.9 KB	Aug 4, 20

When you open R Studio, you will see four panes. On the top-left corner is the Program Pane. This is where you will type and edit your R code in a text file with a .R ending (denoting an R script file). On the bottom-left corner is the Console Pane. This is where code from the Program Pane will be executed. On the top-right corner is the Environment Pane, this is where any assigned objects in the program will appear once they have been executed. Finally, on the Files Pane is located in the bottom-right corner. This is where you can access R help, see which packages have been installed, and take a look at the plots that have been produced.

<u>Creating an R Studio Project</u>

RStudio has features that allow us to work efficiently and manage all the data and outputs associated with each other in an organized way. All data and outputs associated with a certain analysis or manuscript can be assigned to a project.

Let's start by creating a new project in R studio . Click on file, then new project and name a new directory called Lab_1

Before we start using R, it is important to keep the following coding and data management practices in mind:

1)Treat data as read only
#Working with data interactively (e.g., in Excel) where they can be modified means you are never sure of where the data came from, or how it has been modified since collection. It is therefore a good idea to treat your data as “read-only”.

2) Store clean data separately from dirty data
In many cases, data needs to be pre-processed before getting it into an R readable format. Store pre-processed data in a separate file from your project. Cleaned data should be treated as readonly

3) Treat output generated from R as disposable
You should be able to generate all outputs from your R script. Any output you have should be able to be regenerated from your R script

Also, we should get into a habit of organizing out data when we modify datasets and run analyses in R. Good Enough Practices for Scientific Computing gives the following recommendations for project organization:

Each project in its own directory, which is named after the project.

Text documents associated with the project in the doc directory.

Raw data (unprocessed/dity data) and metadata should be in the data directory

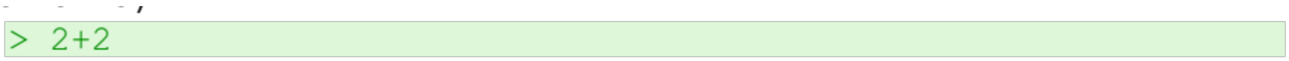
Output generated during analysis should be in a results directory.

R script/source for the project’s scripts and programs in the src directory, and programs brought in from elsewhere or compiled locally in the bin directory.

Once you have created a project, move your cursor to the console and you will see:

```
>
```

This is where commands are entered. R like most other programming languages can perform arithmetic operations. Type 2+2 in the console and press enter:



```
> 2+2
```

Your output should look like this:

```
> 2+2
[1] 4
```

Please open the R script labelled Lab 1 in R Studio. Under File tab, click file and then click Open File. Please go to the directory where you have downloaded or stored Lab1.R and click open. This should load the R script file onto the Program Pane. For the rest of this lab, we will review and edit this script file.

<u>Lab 1 assignment</u>

You will be assessing the data provided in the gapminder.csv file. Please make sure to comment with every function that you use, explaining what the function used does.

1) Please read the gapminder.csv file into R. How many variables does this dataset have? What are the different classes of data in this dataset?

2) How many countries are included in the dataset? How many continents are represented? Which continents are not included? Why?

3) Which is the oldest time point included in the dataset? Which is the most recent?

4) Which is the most populous country? Which is the least populous country in the dataset? Which country has the greatest number of datapoints? Which country has the least number of datapoints?

5) What is the mean and median life expectancy for each country? What is the mean and median life expectancy for each continent? Which country and continent have the greatest mean life expectancy? Which country and continent have the least mean life expectancy? Which country has the most variation in life expectancy across time (hint: use the sd function)? Which country has the least variation in life expectancy across time?

6) What is the mean and median lgdp per capita for each country? What is the mean and median gdp per capita for each continent? Which country and continent have the greatest gdp per capita? Which country and continent have the least gdp per capita? Which country has the most variation in gdp per capita across time (hint: use the sd function)? Which country has the least variation in gdp per capita across time?

7) Make a scatterplot of life expectancy (y) and gdp per capita (x) with the points coloured by continent. Make sure that the x and y axes are labelled and a legend is included.