CS 251 Statistical Computing

HOP 5: R for statistical project

2/19/2020 Developed by Aya Khalil

04/02/2020 Reviewed by Apiwat Chuaphan

School of Technology and Computing @City University of Seattle (CityU)

**Before You Start**

* If you already finished this module through any CityU Technology Institute (TI) courses,  
  just skim this module and skip it.
* Version numbers may not match with the guide. But that should be fine.  
  If given the option to choose between stable release (long-term support) or most recent, please choose the stable release.
* This guide targets Windows OS users. So, MacOS users may have different commands to input in the shell/terminal.
* We cannot explain every step. **This cookbook always needs your own creative judgement.**
* **For your working directory, use your course number.** The hands-on tutorial may use a different course number as an example.

**Learning Outcomes**

* Read CSV file
* Basic data processing

**Resource**

* Hui, E. G. M. (2019). [*Learn R for applied statistics: With data visualization, regressions, and statistics*](https://login.proxy.cityu.edu/sso/skillport?context=144516). Apress.
* Data Science and Machine Learning BootCamp with R online course

**Reading Data Files**

R programming allow you to import a data set, which can be comma-separated values (CSV) file, Excel file, tab-separated file, JSON file, or others. Reading data into the R

console or R is important, since you must have some data before we can do statistical computing and understand the data.

**Setup Working Environment for Module5**

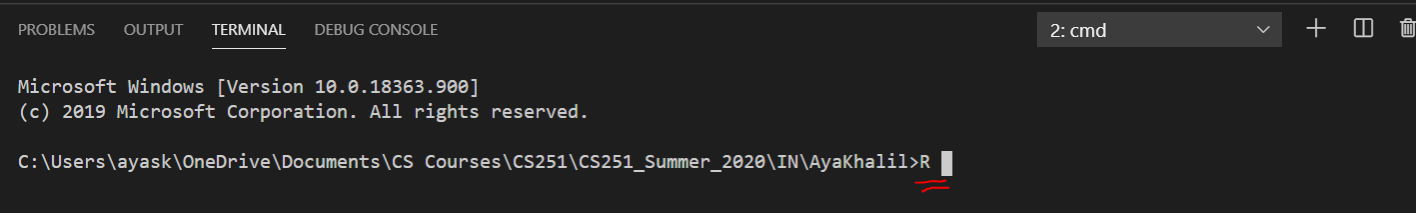
1. Open VS Code.
2. Go to your project folder

* **online student:** Open CS251 \_Fall\_2020/**ON**/FirstnameLastname /. ( File > Open )
* **onsite student:** Open CS251 \_ Fall \_2020/**IN**/FirstnameLastname. ( File > Open )

1. Then create “**Module5**” directory in the VSCode.

>>>mkdir Module5

**To open the R terminal, write in the console R**



**Then enter**

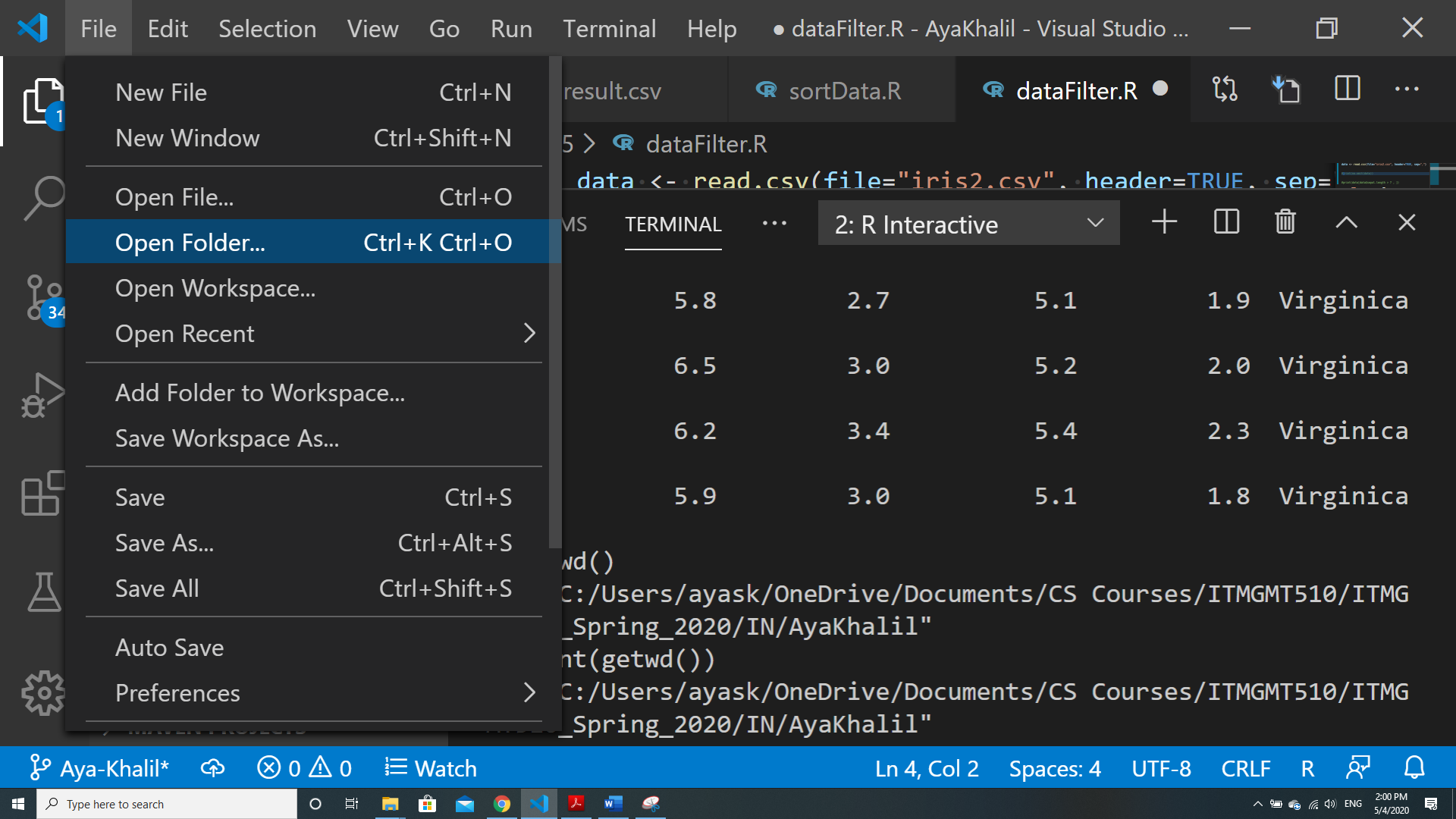
Before we look into importing data into the R console, you must determine your workplace or work directory first. You should always set the current workspace directory to tell R

the location of your current project folder. This allows for easier references to data files and scripts.

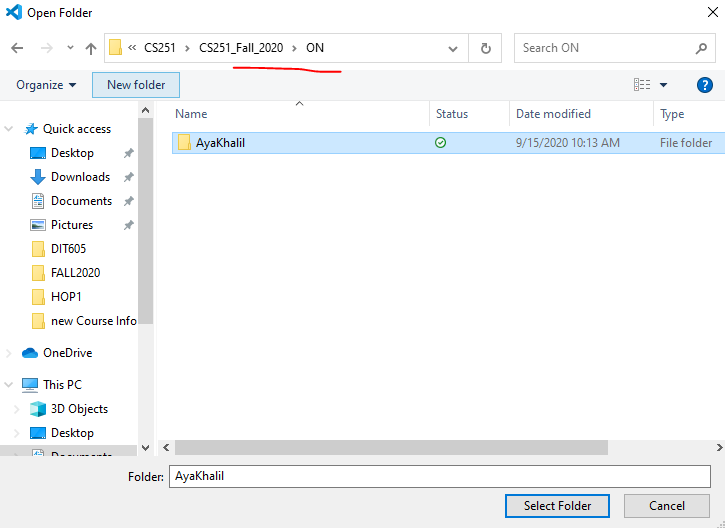
For me, my work directory is " C:\Users\ayask\OneDrive\Documents\CS Courses\CS251\CS251\_ Fall \_2020\IN\AyaKhalil”

**So, to make sure that your work directory is yourname folder**

Go to file> open folder



And choose your Name folder

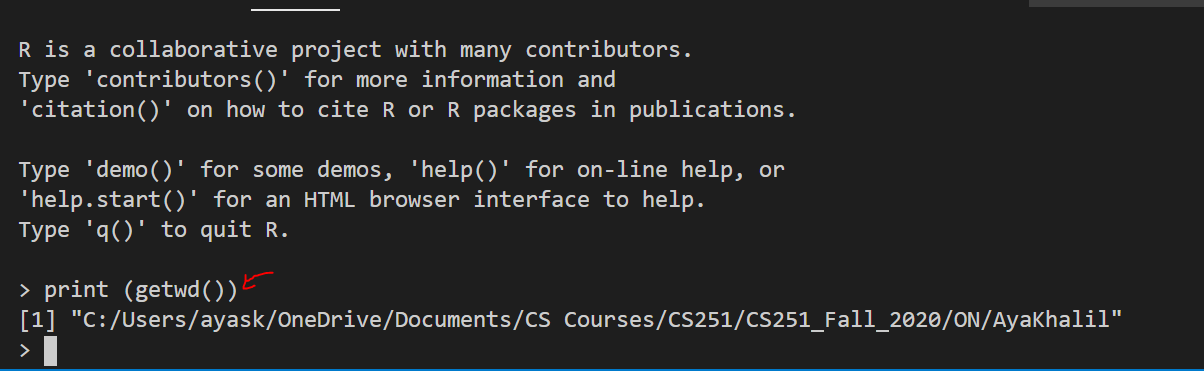


To print the current work directory, you use the getwd() function:

-Open the R console and type print(getwd())

**-For mac user, type in the R console**

**>getwd()**



You can set the work directory using the setwd() function:

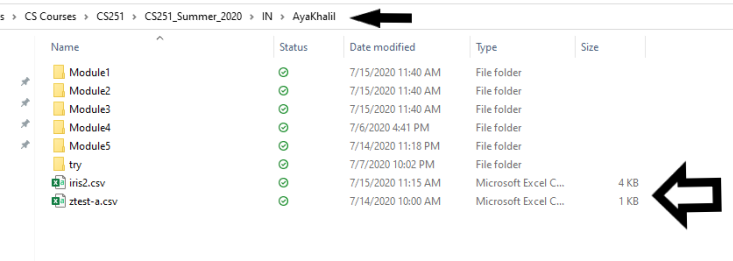
#set the current workspace location

**For example,**

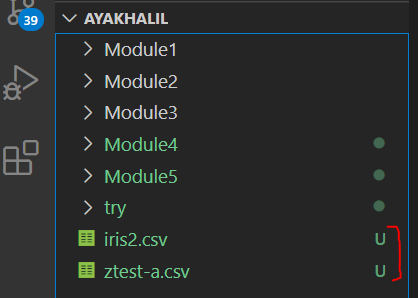
setwd("D:/R");

**-Download the attached CSV files “iris2.csv” & “ztest-a.csv” and copy it to your name folder.**

* **Copy the file to your name folder because this is the working directory.**



**-Make sure you have the iris2.csv & ztest-a.csv files under yourname folder**

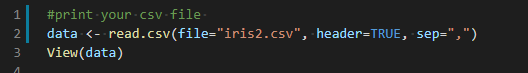


**You should be in:**

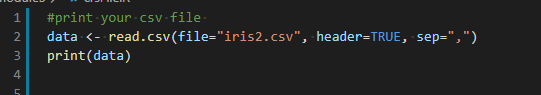
* **online student:** CS251 \_Summer\_2020/**ON**/FirstnameLastname
* **onsite student:** CS2511 \_Summer\_2020/**IN**/FirstnameLastname

-**In Module5 project folder, create new file csvFile.R**

- Type the following code in csvFile.R file (Mac users need to install XQuartz from <https://www.xquartz.org> to run View function)



**Another solution to see the output: If View does not work you can use print(data)**



Save your code, file>save or ctrl+s

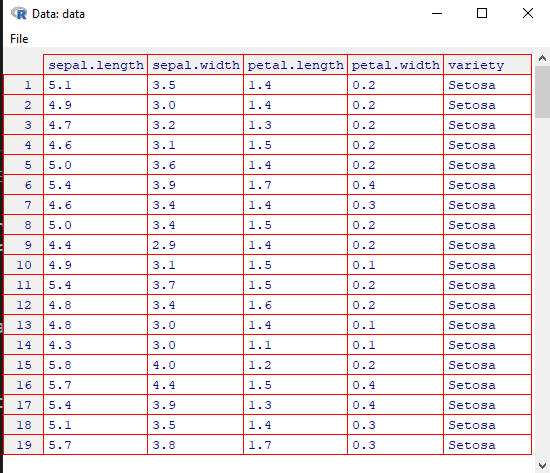
-Run your code:

- Select Run Source button:

-When you select the button for first time, it will open the R terminal

- Select the same button one more time to run and display the code.

**Output:**



**Explanation:**

The function you use to read a .csv file is

> data <- read.csv(file="data.csv", header=TRUE, sep=",");

file is the name or the data file path that you are going to read. header is a logical value to determine whether the names of the variables are in the first line. sep is the separator

character, and quote is the quoting characters with "\"" for " and "\'" for '.

**Basic Data Processing**

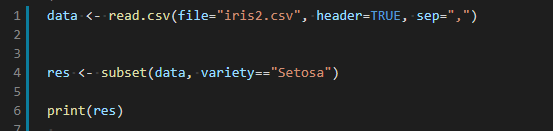
You may need to do some simple data processing like selecting data, sorting data, filtering data, getting unique values, and removing missing values.

**You should be in:**

* **online student:** CS251 \_Fall\_2020/**ON**/FirstnameLastname
* **onsite student:** CS2511 \_ Fall \_2020/**IN**/FirstnameLastname

-**In Module5 project folder, create new file selectData.R**

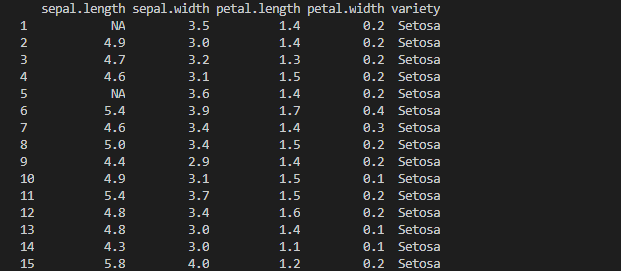
- Type the following code in selectData.R file



-Save your code, file>save or ctrl+s

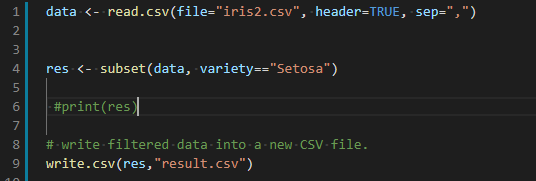
-Run your code:Select Run Source button

-Output: It will print all the Setosa

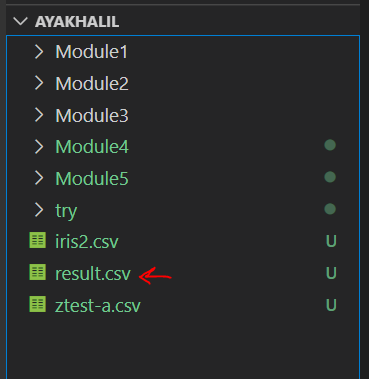


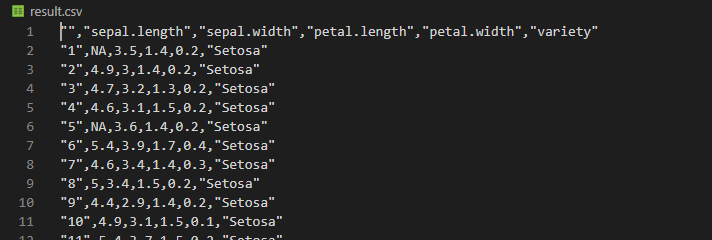
**-We can print the output in another csv file**

-Type the following to edit the previous code



-Save & Run

-o/p 



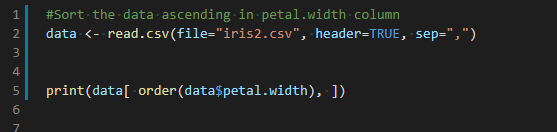
**Data Sorting:**

**You should be in:**

* **online student:** CS251 \_ Fall \_2020/**ON**/FirstnameLastname
* **onsite student:** CS2511 \_ Fall \_2020/**IN**/FirstnameLastname

**-In Module5 project folder, create new file sortData.R**

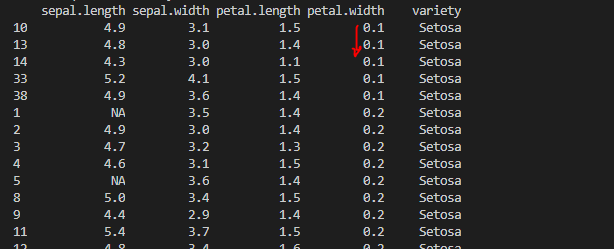
- Type the following code in sortData.R file



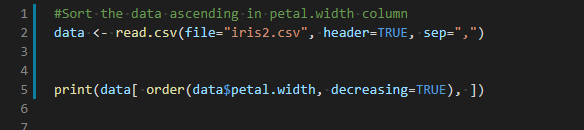
-Save your code, file>save or ctrl+s

-Run your code: Select Run Source button

-Output:



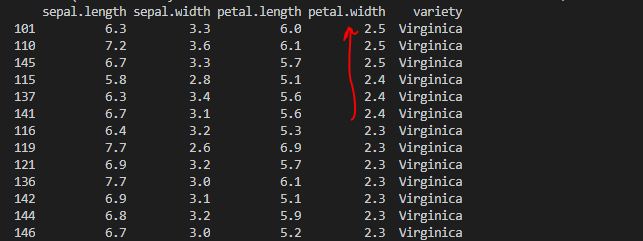
* you can sort the data in descending order
* Type the following to update sortData.R file



-Save your code, file>save or ctrl+s

-Run your code: Select Run Source button

-Output:



**Data Filtering**

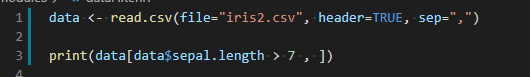
You can filter the data using Boolean expressions and statements

**You should be in:**

* **online student:** CS251 \_ Fall \_2020/**ON**/FirstnameLastname
* **onsite student:** CS2511 \_ Fall \_2020/**IN**/FirstnameLastname

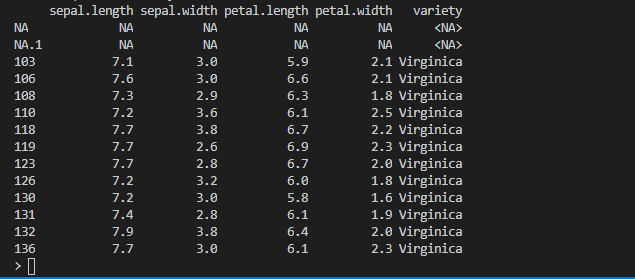
**In Module5 project folder, create new file dataFilter.R**

Type the following code in dataFilter.R file



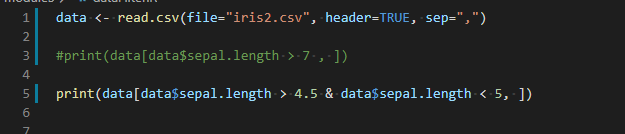
* Save your code, file>save or ctrl+s
* Run your code: Select Run Source button

Output

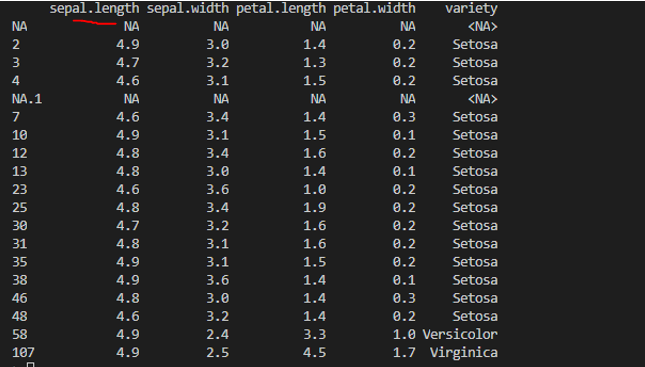


**You can also filter the data with more complex expressions**

**Type the following to update dataFilter.R file**



Run your code: Select Run Source button



**Notice that the values of sepal length columns is between 4.5 & 5**

**Removing Duplicates**

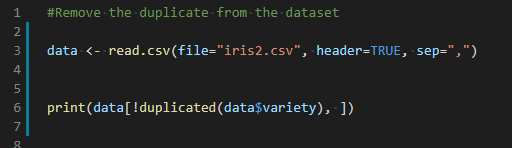
You can remove duplicates based on the x variable.

**You should be in:**

* **online student:** CS251 \_ Fall \_2020/**ON**/FirstnameLastname
* **onsite student:** CS2511 \_ Fall \_2020/**IN**/FirstnameLastname

**In Module5 project folder, create new file removeDup.R**

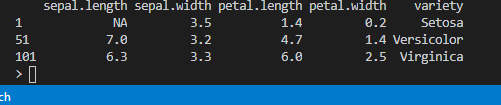
- Type the following code in removeDup.R file



-Save your code, file>save or ctrl+s

-Run your code: Select Run Source button

-Output:



**Removing NA values**

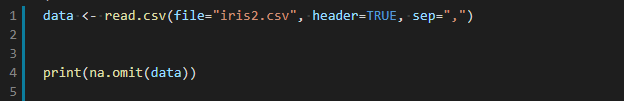
**NA stands for Not Available, so we can delete not available variables**

**You should be in:**

* **online student:** CS251 \_ Fall \_2020/**ON**/FirstnameLastname
* **onsite student:** CS2511 \_ Fall \_2020/**IN**/FirstnameLastname

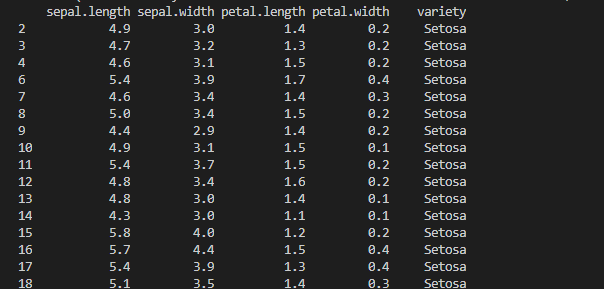
**In Module5**, create removeNA.R

Type the following in removeNA.R



-Save & Run

-o/p: Notice that no NA any more



**(MAC Users)**

**-You don’t have to save image for the workspace commands**

