

Basics of Data Science using R

Hira Anees Awan

Research Intern - Tomaras Lab

Master of Biostatistics, Class of 2021

Learning Objectives

At the conclusion of this workshop, you will be able:

- To load text files and csv files into R
- To understand and use different data types in R widely used in Data Science
- To use an R library
- To manipulate data using powerful R libraries
- To calculate descriptive statistics on different kinds of features in a data
- To visualize different features of a dataset using a variety of plots
- To analyze data using an R library (I am not primarily focusing on data analysis part because there is a huge variety of algorithms available for different kinds of data and the type of question you are trying to answer using that data.)
- Bonus: Neural Networks in R, Principal Component Analysis in R

Pre-requisite: R & R Studio






1. Download R from <http://cran.us.r-project.org/> .
2. Click on Download R for you specific OS. Click on base. Click on Download R 3.3.2 for whatever OS you are using (or a newer version that appears).
3. Install R. Leave all default settings in the installation options.
4. Download RStudio Desktop for your specific OS from <http://rstudio.org/download/desktop>. Choose default installation options.
5. Open RStudio.

In case you need a video tutorial to carry out the steps above, you can visit this [link](#).

Steps

Once you are done with installing R and R studio on your laptop follow the following steps:

- Create a Folder on your Desktop (for the ease of access) and name it 'RWorkshop_TomarasLab'.
- Place RWorkshop.Rmd file, RWorkshop.nb.html file, train.csv, test.csv, and Tab_Delimited_Text_File.txt file in that folder.

Name	Status	Date modified
 RWorkshop.nb	✓	4/9/2020 3:46 PM
 RWorkshop	✓	4/9/2020 9:23 PM
 Tab_Delimited_Text_File	✓	4/5/2020 11:32 PM
 test	✓	3/29/2020 9:38 PM
 train	✓	3/29/2020 9:32 PM

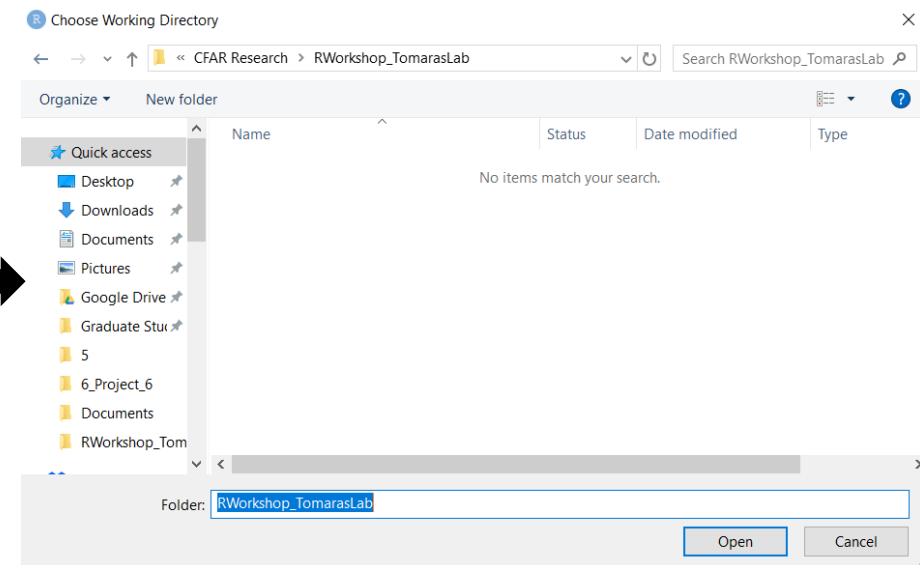
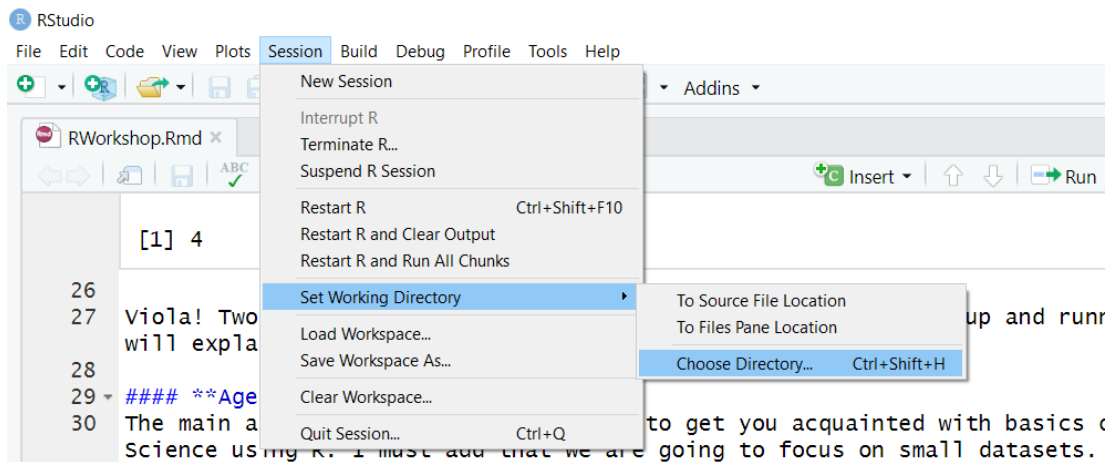
***Do not forget the path of this folder. You will need this path for later use.*

Continued

Open R studio and perform the following steps to set the current working directory.

Benefit: You will not have to use complete path names whenever you want to access a data file.

- In the top bar, click on 'Session', go to 'Set Working Directory' > 'Choose Directory...'. This will prompt you to give the path of the folder where all your files live. Give the path of the folder called 'RWorkshop_TomasasLab'. After doing so, open Rworkshop.Rmd file using RStudio.



Continued...

- Now, locate the console and paste the given code in the console and hit enter. I will not be using all these libraries in today's tutorial, but [this](#) book has some excellent examples that you can look at and explore these libraries.
- In case, you want a general method to install a package in R, use one of the following steps.
 - ...in RStudio click Tools -> Install packages...
 - in the console run: `install.packages("plotly")`
- Loading an R package is easy. Write the following line of code before using the package:
 - `library("plotly")`

```
pkgs <- c("ggplot2", "dplyr", "tidyr",  
  "mosaicData", "carData",  
  "VIM", "scales", "treemapify",  
  "gapminder", "ggmap", "choroplethr",  
  "choroplethrMaps", "CGPfunctions",  
  "ggcorrplot", "visreg",  
  "gcookbook", "forcats",  
  "survival", "survminer",  
  "ggalluvial", "ggridges",  
  "GGally", "superheat",  
  "waterfalls", "factoextra",  
  "networkD3", "ggthemes",  
  "hrbrthemes", "ggpol", "neuralnet",  
  "ggbeeswarm" )
```

```
install.packages(pkgs)
```

Let's begin...

- For the rest of the workshop, we will be following Rworkshop.Rmd file.

Some comments about the dataset:

- I downloaded the given dataset from Kaggle.
- You can see the details of this dataset [here](#).
- I will be using the train and test data only for this tutorial. Test data does not contain survival variable.

Variable	Definition	Key
PassengerId	Passenger Id	
Survived	Survival	0=No, 1=Yes
pclass	Ticket class	
Name	Name	
sex	sex	
Age	Age in years	
sibsp	# of siblings/spouses aboard	
parch	# of parents/children aboard	
Ticket	Ticket number	
Fare	Passenger fare	
cabin	Cabin number	
embarked	Port of embarkation	C= Cherbourg, Q = Queenstown, S = Southampton