Business Statistics

A Guide for BUAD 231

J. Alejandro Gelves

12/27/22

Table of contents

List of Figures

List of Tables

Introduction

"Whatever you would make habitual, practice it; and if you would not make a thing habitual, do not practice it, but accustom yourself to something else." *Epictetus*

How often do we feel bad about ourselves because we procrastinated, squandered our time, or did not accomplish something meaningful during the day? Making the right decisions takes practice. In this book, I invite you to practice the skills you have learned in BUAD 231 and the skills of focus, dedication, and consistency. Choose a day in the week and start by dedicating some fixed time to these problems (e.g., 15-30 minutes). The idea is to work on consistency (i.e., returning to the book weekly for a given amount of time). Some of us will find that concentrating is challenging. Your next task is to reduce distractions (i.e., the phone, t.v. or even your thoughts about the future). If you keep trying and returning to the book, you will improve at Business Statistics and learn to study with focus and consistency. All it takes is practice. Remember, you are what you practice!

The problems in this book are designed to help you master statistics and its application in R. I recommend reviewing Grolemund (2014) if you need additional help learning R. Finally, I have provided a list of concepts at the beginning of every chapter. Enjoy!

Why R?

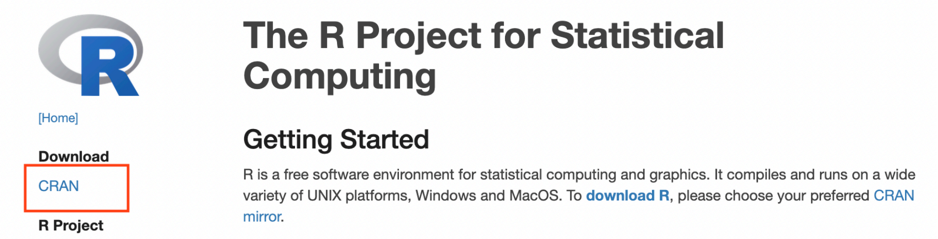
We will be using R to apply the lessons we learn in BUAD 231. R is a language and environment for statistical computing and graphics. There are several advantages to using the R software for statistical analysis and data science. Some of the main benefits include:

- R is a **powerful and flexible programming language** that allows users to manipulate and analyze data in many different ways.
- R has a large and **active community of users**, who have developed a wide range of packages and tools for data analysis and visualization.
- R is **free and open-source**, which makes it accessible to anyone who wants to use it.
- R is **widely used** in academia and industry, which means that there are many resources and tutorials available to help users learn how to use it.
- R is well-suited for working with **large and complex datasets**, and it can handle data from many different sources.
- R can be **easily integrated** with other tools and software, such as databases, visualization tools, and machine learning algorithms.

Overall, R is a powerful and versatile tool for data analysis and data science, and it offers many benefits to users who want to work with data.

Installing R

To install R, visit the R webpage at https://www.r-project.org/. Once in the website, click on the CRAN hyperlink.

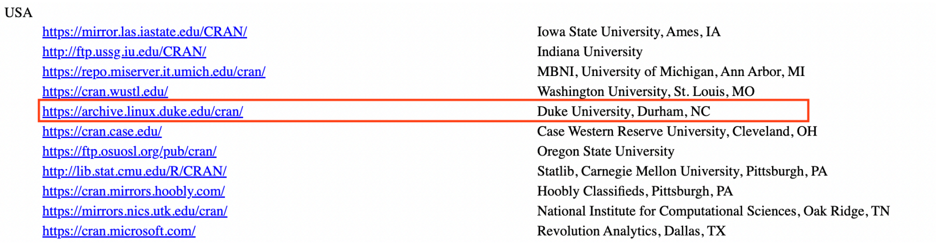


The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

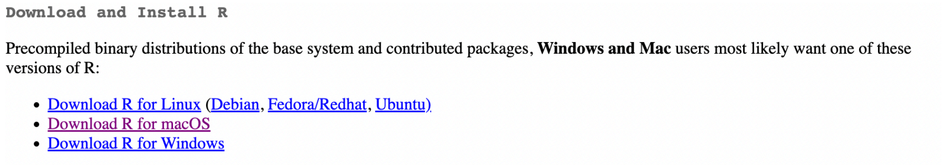
Here you can select the CRAN mirror. Scroll down until you see USA. You are free to choose any mirror you like, I recommend using the Duke University mirror.



USA

https://mirror.las.iastate.edu/CRAN/	Iowa State University, Ames, IA
http://ftp.ussg.iu.edu/CRAN/	Indiana University
https://repo.miserver.it.umich.edu/cran/	MBNI, University of Michigan, Ann Arbor, MI
https://cran.wustl.edu/	Washington University, St. Louis, MO
https://archive.linux.duke.edu/cran/	Duke University, Durham, NC
https://cran.case.edu/	Case Western Reserve University, Cleveland, OH
https://ftp.osuosl.org/pub/cran/	Oregon State University
http://lib.stat.cmu.edu/R/CRAN/	Statlib, Carnegie Mellon University, Pittsburgh, PA
https://cran.mirrors.hoobly.com/	Hoobly Classifieds, Pittsburgh, PA
https://mirrors.nics.utk.edu/cran/	National Institute for Computational Sciences, Oak Ridge, TN
https://cran.microsoft.com/	Revolution Analytics, Dallas, TX

Once you click on the hyperlink, you will be prompted to choose the download for your operating system. Depending on your operating system, choose either a Windows or Macintosh download.



Download and Install R

Precompiled binary distributions of the base system and contributed packages, **Windows and Mac** users most likely want one of these versions of R:

- Download R for Linux (Debian, Fedora/Redhat, Ubuntu)
- Download R for macOS
- Download R for Windows

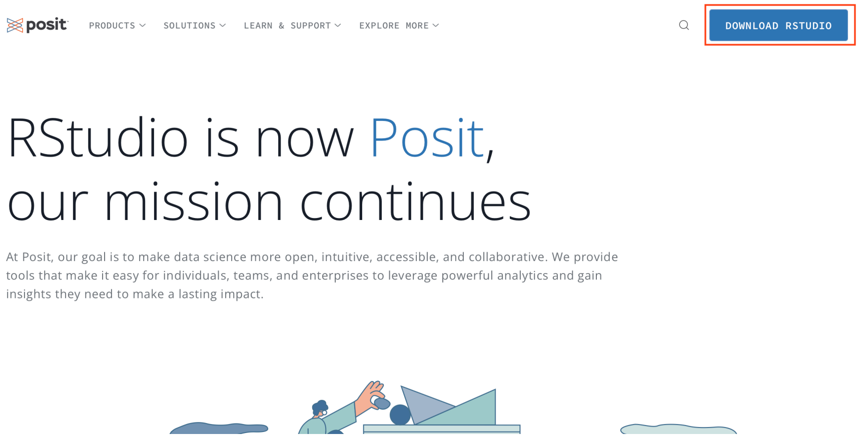
Follow all prompts and complete installation.

Installing RStudio

Visit the Posit website at https://posit.co. Once on the website, hover to the top right of the screen. You will see a “Download RStudio” blue button.

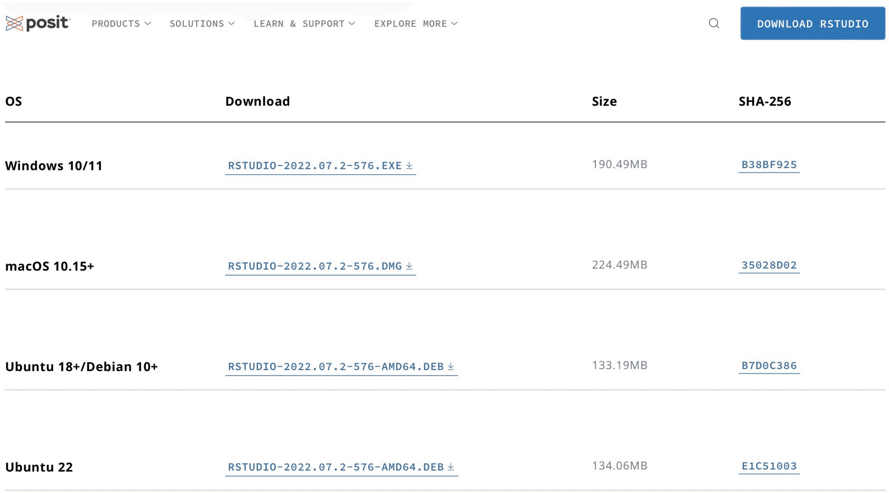
Next, scroll down until you reach the RStudio desktop section. Click once more on “Download RStudio”. You can now just jump to Step 2 since you have already downloaded R. Finally, choose the desired download depending on your operating system.

It is important to note that RStudio will not work if R is not installed. You can think of R as the engine and RStudio as the interface.



RStudio is now Posit,
our mission continues

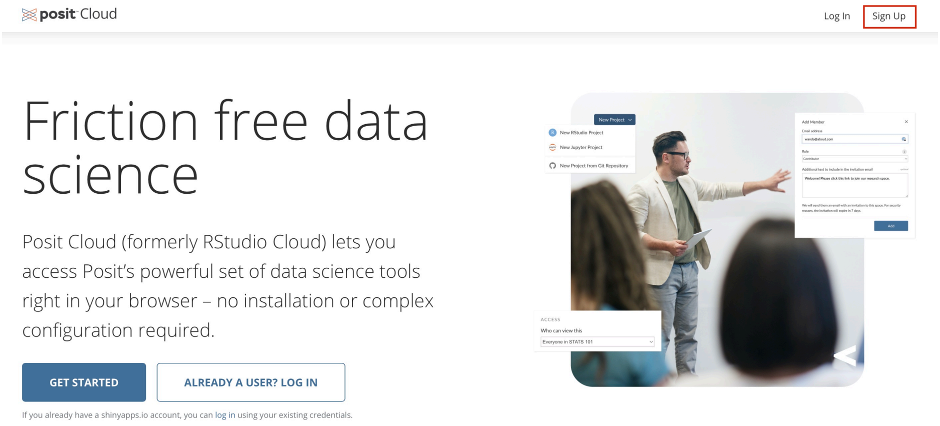
At Posit, our goal is to make data science more open, intuitive, accessible, and collaborative. We provide tools that make it easy for individuals, teams, and enterprises to leverage powerful analytics and gain insights they need to make a lasting impact.



OS	Download	Size	SHA-256
Windows 10/11	RSTUDIO-2022.07.2-576.EXE	190.49MB	B38BF925
macOS 10.15+	RSTUDIO-2022.07.2-576.DMG	224.49MB	35028D02
Ubuntu 18+/Debian 10+	RSTUDIO-2022.07.2-576-AMD64.DEB	133.19MB	B7D0C386
Ubuntu 22	RSTUDIO-2022.07.2-576-AMD64.DEB	134.06MB	E1C51003

Posit Cloud

If you do not wish to install R, you can always use the cloud version. To do this, visit https://posit.cloud/. On the main page click on the "Sign Up" button.



Friction free data science

Posit Cloud (formerly RStudio Cloud) lets you access Posit's powerful set of data science tools right in your browser – no installation or complex configuration required.

If you already have a shinyapps.io account, you can log in using your existing credentials.

Email address: wanda@about.com

Role: Contributor

Welcome! Please click this link to join our research space.

We will send them an email with an invitation to this space. For security reasons, the invitation will expire in 7 days.

Who can view this: Everyone in STATS 101

Choose the "Cloud Free" option and log in using your Google credentials (if you have a Google account) or sign up if you want to create a new account.

Part I

Descriptive Statistics

1 Descriptive Stats I

1.1 Concepts

Data and Types of Data

Data are facts and figures collected, analyzed and summarized for presentation and interpretation. Data can be classified as:

- **Cross Sectional Data** refers to data collected at the same (or approximately the same) point in time. Ex: NFL standings in 1980 or Country GDP in 2015.
- **Time Series Data** refers to data collected over several time periods. Ex: U.S. inflation rate from 2000-2010 or Tesla deliveries from 2016-2022.
- **Structured Data** resides in a predefined row-column format (tidy).
- **Unstructured Data** do not conform to a pre-defined row-column format. Ex: Text, video, and other multimedia.

Data Sets, Variables and Scales of Measurement

A **data set** contains all data collected for a particular study. Data sets are composed of:

- **Elements** are the entities on which data are collected. Ex: Football teams, countries, and individuals.
- **Observations** are the set of measurements obtained for a particular element.
- **Variables** are a set of characteristics collected for each element.

The **scales of measurements** determine the amount and type of information contained in each variable. In general, variables can be classified as **categorical** or **numerical**.

- **Categorical** (qualitative) data includes labels or names to identify an attribute of each element. Categorical data can be **nominal** or **ordinal**.
  - With **nominal** data, the order of the categories is arbitrary. Ex: Marital Status, Race/Ethnicity, or NFL division.
  - With **ordinal** data, the order or rank of the categories is meaningful. Ex: Rating, Difficulty Level, or Spice Level.
- **Numerical** (quantitative) include numerical values that indicate how many (discrete) or how much (continuous). The data can be either **interval** or **ratio**.

  - With **interval** data, the distance between values is expressed in terms of a fixed unit of measure. The zero value is arbitrary and does not represent the absence of the characteristic. Ratios are not meaningful. Ex: Temperature or Dates.
  - With **ratio** data, the ratio between values is meaningful. The zero value is not arbitrary and represents the absence of the characteristic. Ex: Prices, Profits, Wins.

Useful R Functions

Base R has some important functions that are helpful when dealing with data. Below is a list that might come handy.

- The `na.omit()` function removes any observations that have a missing value (NA). The resulting data frame has only complete cases.
- The `nrow()` and `ncol()` functions return the number of rows and columns respectively from a data frame.
- The `is.na()` function returns a vector of *True* and *False* that specify if an entry is missing (NA) or not.
- The `summary()` function returns a collection of descriptive statistics from a data frame (or vector). The function also returns whether there are any missing values (NA) in a variable.
- The `as.integer()`, `as.factor()`, `as.double()`, are functions used to coerce your data into a different scale of measurement.

The `dplyr` package has a collection of functions that are useful for data manipulation and transformation. If you are interested in this package you can refer to Wickham (2017). To install, run the following command in the console `install.packages("dplyr")`.

- The `arrange()` function allows you to sort data frames in ascending order. Pair with the `desc()` function to sort the data in descending order.
- The `filter()` function allows you to subset the rows of your data based on a condition.
- The `select()` function allows you to select a subset of variables from your data frame.

1.2 Exercises

The following exercises will help you test your knowledge on the Scales of Measurement. They will also allow you to practice some basic data "wrangling" in R. In these exercises you will:

- Identify numerical and categorical data.
- Classify data according to their scale of measurement.
- Sort and filter data in R.
- Handle missing values (NA's) in R.

Answers are provided below. Try not to peak until you have a formulated your own answer and double checked your work for any mistakes.