Setting Up R for Analytics

Why this document?

The general aim of this document is to help you, a student in an analytics-focused course, learn to use the software platform R as part of your learning of statistics, operational research, and data analytics that accompanies nearly every domain of knowledge, from epidemiology to financial engineering. I will add new material as we cover it in class and edit old material based on feedback from you to make it clearer for you and future students. I suggest that you do not print new versions of this and other support documents. Instead please replace your electronic copy from time to time. Good luck as you begin your quest to master introductory concepts and their application!

The specific aim of this document is to show you how to install R an integrated development environment (IDE), RStudio, and a documentation system R Markdown on your personal computing platform (also known as your personal computer). This will enable you to learn the statistical concepts usually included in an analytics course with explanations and examples aimed at the appropriate level. This document purposely does not attempt to teach you about R's many fundamental and advanced features.

What is R?

R is software for interacting with data along a variety of user generated paths. With R you can create sophisticated (even interactive) graphs, you can carry out statistical and operational research analyses, and you can create and run simulations. R is also a programming language with an extensive set of built-in functions. With increasing experience, you can extend the language and write your own code to build your own financial analytical tools. Advanced users can even incorporate functions written in other languages, such as C, C++, and Fortran.

The current version of R derives from the S language. S has been around for more than twenty years and has been with extrensive use in statistics and finance, first as S and then as the commercially available S-PLUS. R is an open source implementation of the S language that is now a viable alternative to S-PLUS. A core team of statisticians and many other contributors work to update and improve R and to make versions that run well under all of the most popular operating systems. Importantly, R is a free, high-quality statistical software that will be useful as you learn financial analytics even though it is also a first-rate tool for professional statisticians, operational researchers, and financial analysts and engineers.

R for analytics

There are several reasons that make R an excellent choice of software for an analytics course. Some benefits of using R include:

- R is free and available online. R is open-source and runs on UNIX, Windows, and Macintosh operating systems.
- R has a well-documented, context-based, help system enhanced by a wide, and deep, ranging user community globally and across several disciplines.
- R has excellent native static graphing capabilities. Interactive dynamic graphics are evolving along with the ability to embed analytics into online applications. With R you can build dashboards and websites to communicate results dynamically with consumers of the analytics you generate.

- Practitioners can easily migrate to the commercially supported `S-Plus` program, if commercial software is required. `S` and `S-Plus` are the immediate ancestors of the R programming environment. Cloud computing is now available with large data implementations.
- `R`'s language has a powerful, easy-to-learn syntax with many built-in statistical and operational research functions. Just as important are the extensive web-scraping, text structuring, object class construction, and the extensible functional programming aspects of the language. A formal language definition is being developed. This will yield more standardization and better control of the language in future versions.
- `R` is a computer programming language. For programmers it will feel more familiar than for others, for example Excel users. `R` requires array thinking and object relationships that are not necessarily native, but indeed are possible, in an Excel spreadsheet environment. In many ways, the Excel style and `R` style of environments complement one another.
- Even though it is not necessarily the simplest software to use, the basics are easy enough to master, so that learning to use `R` need not interfere with learning the statistical, operational research, data, and domain-specific concepts encountered in an analytics-focused course.

There is at least one drawback.

- The primary hurdle to using `R` is that most existing documentation and plethora of packages are written for an audience that is knowledgable about statistics and operational research and has experience with other statistical computing programs. In contrast, this course intends to make `R` accessible to you, especially those who are new to both statistical concepts and statistical computing.

Some useful `R` resources

There are many `R` books useful for managing implementation of models in this course. Three useful `R` books include:

1. Paul Teetor, *The R Cookbook*
2. Phil Spector, *Data Manipulation with R*
3. Norman Matloff, *The Art of R Programming: A Tour of Statistical Software Design*
4. John Taveras, *R for Excel Users* at https://www.rforexcelusers.com/book/.

The first one will serve as our `R` textbook. The other books are extremely valuable reference works. You will ultimately need all three (and whatever else you can get your hands on) in your professional work. John Taveras's book is an excellent bridge and compendium of Excel and R practices.

Much is available in books, e-books, and online for free. This is an extensive online user community that links expert and novice modelers globally.

1. The standard start-up is at CRAN http://cran.r-project.org/manuals.html. A script in the appendix can be dropped into a workspace and played with easily.
2. Julian Faraway's https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf is a fairly complete course on regression where you can imbibe deeply of the many ways to use `R` in statistics.
3. Along econometrics lines is Grant Farnsworth's https://cran.r-project.org/doc/contrib/Farnsworth-EconometricsInR.pdf.
4. Winston Chang's http://www.cookbook-r.com/ and Hadley Wickham's example at http://ggplot2.org/ are online graphics resources.
5. Stack Overflow is a programming user community with an `R` thread at http://stackoverflow.com/questions/tagged/r. The odds are that if you have a problem, error, or question, it has already been asked, and answered, on this site.

6. For using `R Markdown` there is a short reference at https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf. Cosma Shalizi has a much more extensive manual at http://www.stat.cmu.edu/~cshalizi/rmarkdown/.

R on your computer

Directions exist at the `R` website, <(http://cran.rproject.org/)> for installing `R`. There are several `twotorials`, including some on installation that can be helpful at http://www.twotorials.com/.

Here are more explicit instructions that tell you what to do.

Download the software from the CRAN website. There is only one file that you need to obtain (a different file depending on the operating system). Running this file begins the installation process which is straight-forward in most, if not all, systems.

- Download `R` from the web. Go the `R` home page at http://cran.us.r-project.org/.
- If you have **Windows** (95 or later), then perform these actions. Click on the link `Windows (95 and later)`, then click on the link called `base`, and finally click on `rw1071.exe` (or the most recent version which could have a larger number in the file name). This begins the download of a file whose size is currently about 20MB. After the download is complete, double-click on the downloaded file and follow the on screen installation instructions.
- If you have **Macintosh** (OS X), then perform these actions. Click on the link `MacOS (System 8.6 to 9.1 and MacOS X)`, then click on `rm171.sit` (or the most recent version which could have a larger number in the file name) which begins the download. When given a choice to unstuff or save, choose save and save it on your desktop. Double-click on the downloaded file. Your Mac will unstuff the downloaded file and create an `R` folder. Inside this folder, there are many files including one with the `R` logo. You may drag a copy of this to your panel and then drag the whole `R` folder to your `Applications` folder (located on the hard drive). After completing this, you can drag the original downloaded file to your trash bin.

Install `RStudio`

Every software platform has a graphical user interface ("GUI" for short). One of the more popular GUIs, and the one used exclusively in this course, is provided by `RStudio` at http://www.rstudio.com. `RStudio` is a freely distributed integrated development environment (IDE) for `R`. It includes a console to execute code, a syntax-highlighting editor that supports direct code execution, as well as tools for plotting, reviewing code history, debugging code, and managing workspaces. In the following steps you will navigate to the `RStudio` website where you can download `R` and `RStudio`. These steps assume you have a `Windows` or `Mac OSX` operating system.

1. Click on https://www.rstudio.com/products/RStudio/ and navigate down to the `Download Desktop` button and click.
2. Click on the `Download` button for the `RStudio Desktop Personal License` choice.
3. Navigate to the sentence: "RStudio requires R 2.11.1+. If you don't already have R, download it *here*." If you have not downloaded `R` (or want to again), click on `here`. You will be directed to the https://cran.rstudio.com/ website in a new browser tab.

- In the `CRAN` site, click on `Download R for Windows`, or `Download R for (MAC) OS X` depending on the computer you use. This action sends you to a new webpage in the site.
- Click on `base`. This action takes you to the download page itself.

4. If you have Windows

- Click on Download R 3.3.2 for Windows (62 megabytes, 32/64 bit) (as of 11/8/2016; other version numbers may appear later than this date). A Windows installer in an over 70 MB R-3.3.2-win.exe file will download through your browser.
- In the Chrome browser, the installation-executable file will reside in a tray at the bottom of the browser. Click on the up arrow to the right of the file name and click Open in the list box. Follow the many instructions and accept default values throughout.
- Use the default Core and 32-Bit files if you have a Windows 32-bit Operating System. You may want to use 64-Bit files if that is your operating system architecture. You can check this out by going to the Control Panel, then System and Security, then System, and look up the System Type:. It may read for example 32-bit Operating System.
- Click Next to accept defaults. Click Next again to accept placing R in the startup menu folder. Click Next again to use the R icon and alter and create registries. At this point the installer extracts files, creates shortcuts, and completes the installation.
- Click Finish to finish.

4. If you have a MAC OS X

- Click on Download R 3.3.2 for MACs (62 megabytes, 32/64 bit) (as of 11/8/2016; other version numbers may appear later than this date). A Windows installer in an over 70 MB R-3.3.2-win.exe file will download through your browser.
- When given a choice to unstuff or save, choose save and save it on your desktop. Double-click on the downloaded file. Your Mac will unstuff the downloaded file and create an R folder. Inside this folder, there are many files including one with the R logo.
- Inside the R folder drag a copy of R logo file to your panel and then drag the whole R folder to your Applications folder (located on the hard drive).

5. Now go back to RStudio browser tab. Click on RStudio 1.0.44 - Windows Vista/7/8/10 or RStudio 1.0.44 - MAC OS X to download RStudio. Executiable files will download. Follow the directions exactly, and similarly, to the ones above.

Install R Markdown

Click on RStudio in your tray or start up menu. Be sure you are connected to the Internet. A console panel will appear. At the console prompt > type

```
install.packages("rmarkdown")
```

- This action will install the RMarkdown package. This package will enable you to construct documentation for your work in the course. Assignments will be documented using RMarkdown for submission to the learning management system.
- This extremely helpful web page, http://rmarkdown.rstudio.com/gallery.html, is a portal to several examples of R Markdown source files that can be loaded into RStudio, modified, and used with other content for your own work.

Install LaTex

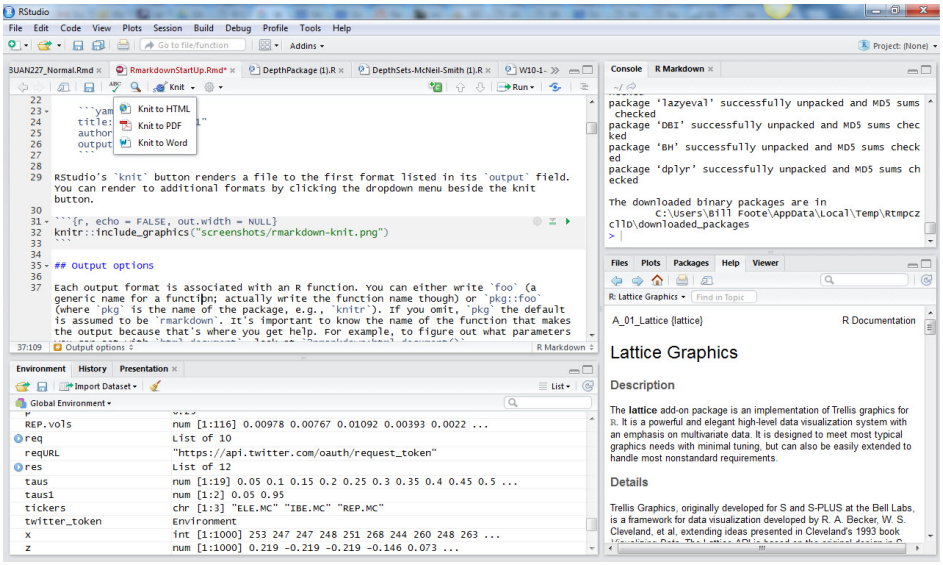
R Markdown uses a text rendering system called LaTeX to render text, including mathematical and graphical content.

1. Install the `MikTeX` document rendering system for `Windows` or `MacTeX` document rendering system for `Mac OS X`.

- For `Windows`, navigate to the https://miktex.org/download page and go to the **64- or 32- bit** installer. Click on the appropriate `Download` button and follow the directions. **Be very sure you select the *COMPLETE* installation**. Frequently Asked Questions (FAQ) can be found at https://docs.miktex.org/faq/. If you have `RStudio` already running, you will have to restart your session.
- For `MAC OS X`, navigate to the http://www.tug.org/mactex/ page and download the `MacTeX` system and follow the directions. This distribution requires Mac OS 10.5 Leopard or higher and runs on Intel or PowerPC processors. **Be very sure you select the *FULL* installation**. Frequently Asked Questions (FAQ) can be found at https://docs.miktex.org/faq/. If you have `RStudio` already running, you will have to restart your session. FAQ can be found at http://www.tug.org/mactex/faq/index.html.

R Markdown

Open `RStudio` and see something like this screenshot...



- You can modify the position and content of the four panes by selecting `View > Panes > Pane Options`.
- Under `File > New File > Rmarkdown` a dialog box invites you to open document, presentation, Shiny, and other files. Upon choosing `documents` you may open up a new file. Under `File > Save As` save the untitle file in an appropriate directory. The `R Markdown` file extension `Rmd` will appear in the file name in your directory.
- When creating a new `Rmarkdown` file, `RStudio` deposits a template that shows you how to use the markdown approach. You can generate a document by clicking on `knit` in the icon ribbon attached to the file name tab in the script pane. If you do not see `knit`, then you might need to install and load the `knitr` package with the following statements in the `R` console. You might need also to restart your RStudio session.

```
install.packages("knitr")
library(knitr)
```

The `Rmd` file contains three types of content:

1. An (optional) **YAML header** surrounded by `---` on the top and the bottom of `YAML` statements. `YAML` is "Yet Another Markdown (or up) Language". Here is an example from this document:

```
---
title: "Setting Up R for Analytics"
author: "Bill Foote"
date: "November 11, 2016"
output: pdf_document
---
```

2. **Chunks** of R code surrounded by "‘ (find this key usually with the ~ symbol).
3. Text mixed with text formatting like `# heading` and `_italics_` and mathematical formulae like `$z = \frac{(\bar x-\mu_0)}{s/\sqrt{n}}$` which will render

$$z = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$$

.

When you open an `.Rmd` file, `RStudio` provides an interface where code, code output, and text documentation are interleaved. You can run each code chunk by clicking the `Run` icon (it looks like a play button at the top of the chunk), or by pressing `Cmd/Ctrl + Shift + Enter`. `RStudio` executes the code and displays the results in the console with the code.

You can write mathematical formulae in an `R Markdown` document as well. For example, here is a formula for net present value.

```
$$
NPV = \sum_{t=0}^{T} \frac{NCF_t}{(1+WACC)^t}
$$
```

This script will render

$$NPV = \sum_{t=0}^{T} \frac{NCF_t}{(1+WACC)^t}$$

- Here are examples of common in file text formatting in `R Markdown`.

```
Text formatting
------------------------------------------------------------

*italic*  or _italic_
**bold**   __bold__
`code`
superscript^2^ and subscript~2~

Headings
------------------------------------------------------------

# 1st Level Header

## 2nd Level Header

### 3rd Level Header

Lists
------------------------------------------------------------

*   Bulleted list item 1
```

```
*   Item 2

    * Item 2a

    * Item 2b

1.  Numbered list item 1

1.  Item 2. The numbers are incremented automatically in the output.

Links and images
------------------------------------------------------------

<http://example.com>

[linked phrase](http://example.com)

![optional caption text](path/to/img.png)

Tables
------------------------------------------------------------

First Header  | Second Header
------------- | -------------
Content Cell  | Content Cell
Content Cell  | Content Cell

Math
------------------------------------------------------------

$\frac{\mu}{\sigma^2}$

\[\frac{\mu}{\sigma^2}]
```

More information will be provided on `R Markdown` documentation throughout the course.

jaRgon

(directly copied from Patrick Burns at http://www.burns-stat.com/documents/tutorials/impatient-r/jargon/, and annotated a bit, for educational use only.)

atomic vector

An object that contains only one form of data. The atomic modes are: *logical*, *numeric*, *complex* and *character*.

attach

The act of adding an item to the search list. You usually attach a package with the `require` function, you attach saved files and objects with the `attach` function.

data frame

A rectangular data object where each column may be a different type of data. Conceptually a generalization of a matrix, but implemented entirely differently.

factor

A data object that represents categorical data. It is possible (and often unfortunate) to confuse a factor with a character vector.

global environment

The first location on the search list, and the place where objects that you create reside. See **search list**.

list

A type of object with possibly multiple components where each component may be an arbitrary object, including a list.

matrix

A rectangular data object where all cells have the same data type. Conceptually a specialization of a data frame, but implemented entirely differently. This object has rows and columns.

package

A collection of `R` objects in a special format that includes help files and such. Most packages primarily or exclusively contain functions, but some packages exclusively contain datasets.

search list

The collection of locations that `R` searches for objects when it is evaluating a command.