Intro to R Statistical Software

UCLA - Econ 221 - Fall 2018

François Geerolf

"Programs must be written for people to read and only incidentally for machines to execute." (Hal Abelson)

Getting started with R Statistical Software

Downloading. You need to install `R` and `Rstudio`:

1. You can get **R statistical software** on the UCLA mirror here. Here is a direct link to "Feather Spray" (R version 3.5.1) for Mac OSX - check first that there does not exist anything more recent. Here is a direct link for Windows.
2. I recommend you use a Graphical User Interface (GNU) for R such as **R Studio**. You can get the latest release here: download here.

Introduction to R. I recommend cheatsheets to get started on R. Many are available, but the 2 main cheatsheets are:
* A Base R Cheatsheet.
* An Advanced R Cheatsheet
Datacamp also have great learning tools for R, as well as Python.

R-markdown

`R-markdown` is a great tool for keeping your workflow organized and keeping track of each one of your research projects: you can add LaTeXvery easily, images, regression tables, graphs, etc. Again, cheatsheets are a good way to learn: there exists a beginner's cheatsheet and a more advanced one. You may also learn from the reference guide here.

Necessary Packages

I will mostly be using `tidyverse`, from Hadley Wickham, for data manipulation as well as plotting data. This cheatsheet has a beginner's introduction to `tidyverse`, and `tidyverse` is presented on this blogpost. `tidyverse` is a powerful collection of R packages that are data tools for transforming and visualizing data. Datacamp has a free tutorial for `tidyverse`, which can get you started. The following packages are particularly useful:

- `dplyr` for data manipulation. Cheatsheet. You will find a tutorial in 4 parts here: Part 1 / Part 2 / Part 3 / Part 4. Note, in particular, the use of pipes %>%:
  - x %>% f(y) is the same as f(x, y).
  - y %>% f(x, ., z) is the same as f(x, y, z).
  - "Piping" with %>% makes code more readable.

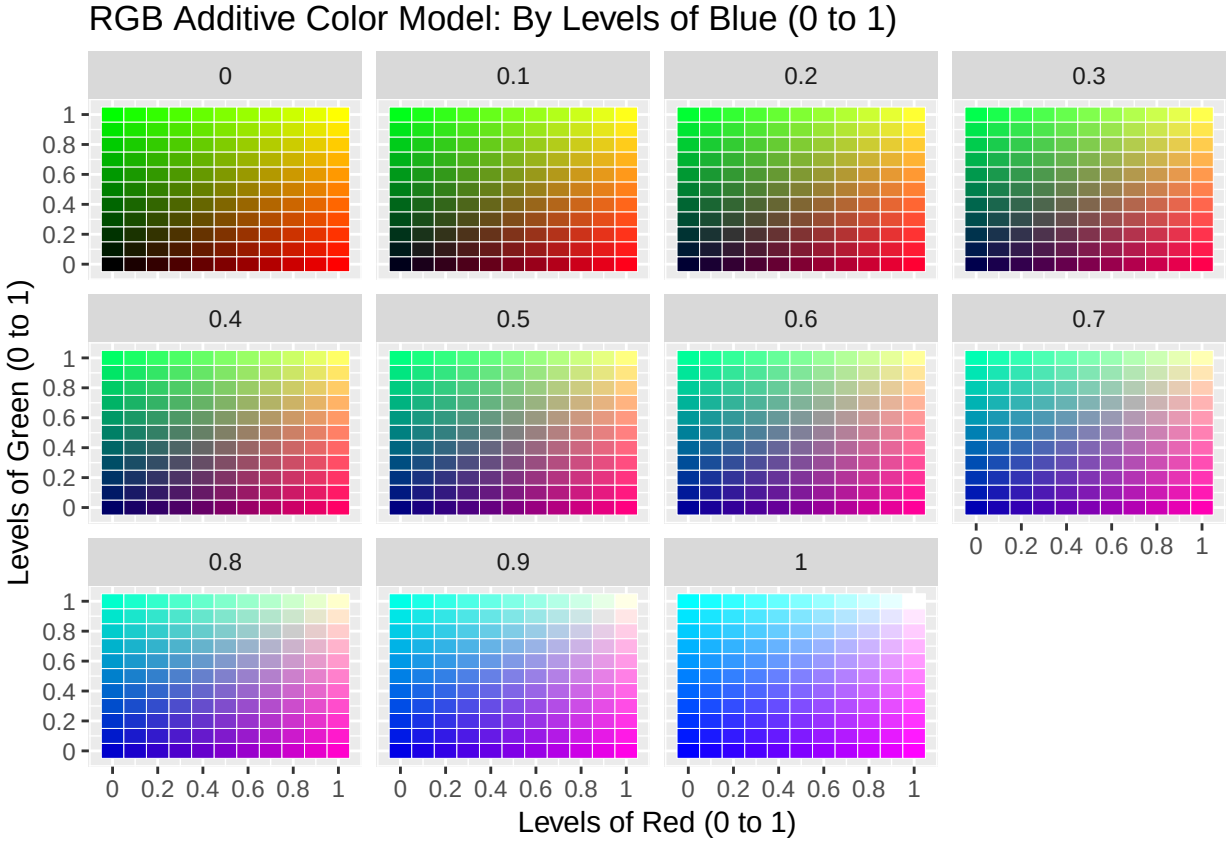
```
iris %>%
  group_by(Species) %>%
  summarise(avg = mean(Sepal.Width)) %>%
  arrange(avg)
```

```
## # A tibble: 3 x 2
##   Species      avg
##   <fct>      <dbl>
```

```
## 1 versicolor  2.77
## 2 virginica   2.97
## 3 setosa      3.43
```

- `ggplot2` for data visualization. Cheatsheet. Combined with `tidyverse`, `ggplot2` proves very powerful. For example, below is a visualization of the RGB additive color model.

```
expand.grid(r = seq(0, 1, 0.1), g = seq(0, 1, 0.1), b = seq(0, 1, 0.1)) %>%
  ggplot() + facet_wrap( ~ b) +
  scale_x_continuous(name = "Levels of Red (0 to 1)",
                     breaks = seq(0.05, 1.05, 0.2),
                     labels = seq (0, 1, 0.2)) +
  scale_y_continuous(name = "Levels of Green (0 to 1)",
                     breaks = seq(0.05, 1.05, 0.2),
                     labels = seq(0, 1, 0.2)) +
  scale_fill_identity() +
  geom_rect(aes(xmin = r, xmax = r + resolution(r),
                ymin = g, ymax = g + resolution(g),
                fill = rgb(r, g, b)),
            color = "white", size = 0.1) +
  ggtitle("RGB Additive Color Model: By Levels of Blue (0 to 1)")
```



- `stringr` for string manipulation. A cheatsheet is available here. If you do want to work on string variables a lot, for example to do web scrapping, then you should learn about regular expressions.
- `readr` to read in data. A cheatsheet is provided here.

In addition to the `tidyverse` collection of R packages, I also use the following packages:

- `lubridate` for working with dates (very useful in macroeconomics !). A cheatsheet is provided here.

Other Packages

Here are other potentially useful packages:

- `tidytext` for analyzing text with the `tidyverse` tools. A great introduction to this package is provided here.

- `bookdown` as a great complement to R-markdown, in order to write more advanced documents. A great introduction is also provided here.