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. First you must get the **R statistical software**, which you may download on the UCLA website here. The latest release (2018-07-02, Feather Spray) is version 3.5.1. For Mac OSX: download here. For Windows: download here.
2. Second, I recommend you use a Graphical User Interface (GNU) for R such as **R Studio**. R Studio's latest release is 1.1.456: download here.

Introduction to R. Cheatsheets are a great way to get started on R. Many are available here, but the 2 main cheatsheets are:

- Base R Cheatsheet.
- Advanced R Cheatsheet

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:

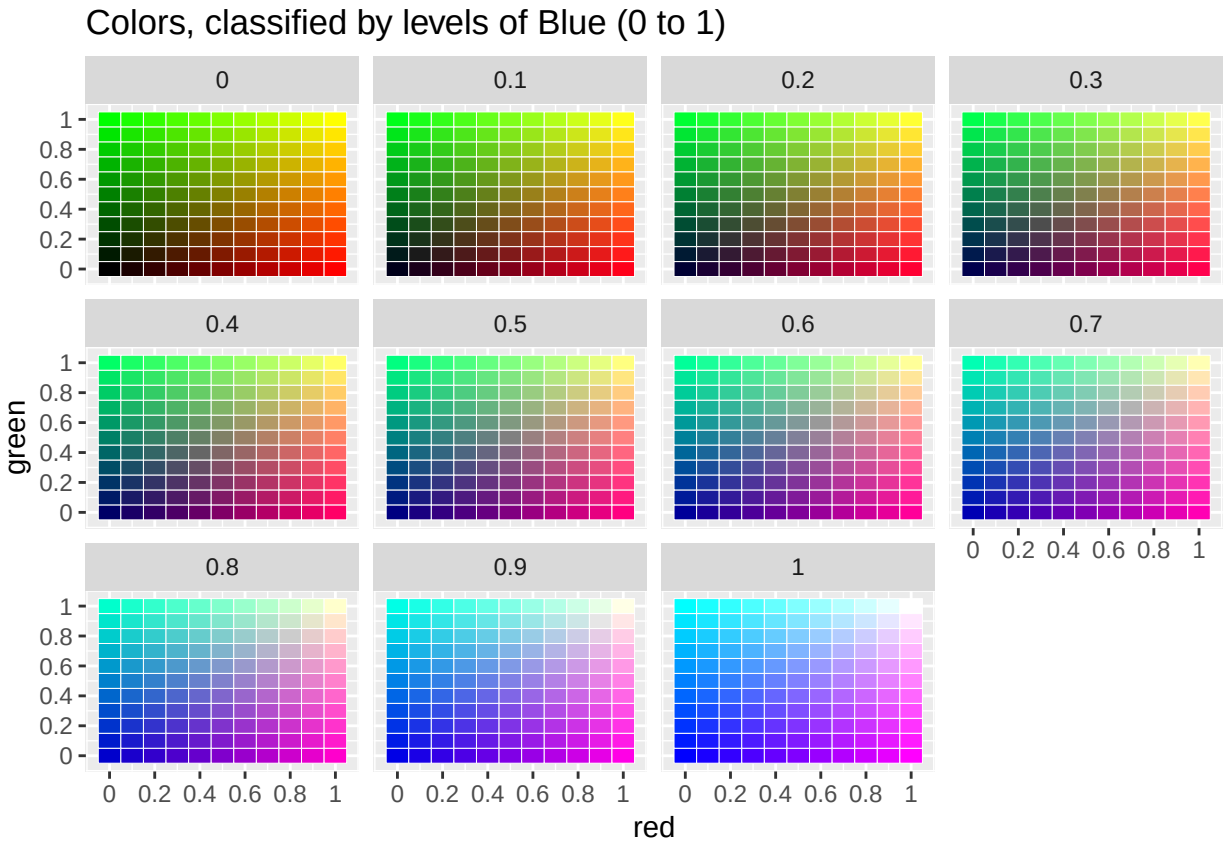
- `dplyr` for data manipulation. Cheatsheet. Note, in particular, the use of pipes %>%:
  - x %>% f(y) is the same as f(x, y).
  - y %>% f(x, ., z) is the same as f(x, y, z).
  - "Piping" with %>% makes code more readable.

```
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 = "red", breaks = seq(0.05, 1.05, 0.2), labels = seq (0, 1, 0.2)) +
  scale_y_continuous(name = "green", 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("Colors, classified by levels of Blue (0 to 1)")
```



- `stringr` for string manipulation. Cheatsheet. Cheatsheet on Regular Expressions.

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

- `lubridate` for working with dates (very useful in macroeconomics !). Cheatsheet.

`tidyverse` also contains `readr` which allows to read in data. Cheatsheet

R-markdown

`R-markdown` is a great tool for keeping your research flow organized and keeping track of each one of your research projects. You can add LaTeXvery easily, regression tables, graphs, etc.

You may start to learn using this cheatsheet, as well as this reference guide.

The learning curve is quite steep but in my opinion, it is really worth it !