Applied Mathematics in Industry from a Data Scientist's Perspective

One or Two Things I Wish I Had Learned In School

Jay Lee
2021/02/10

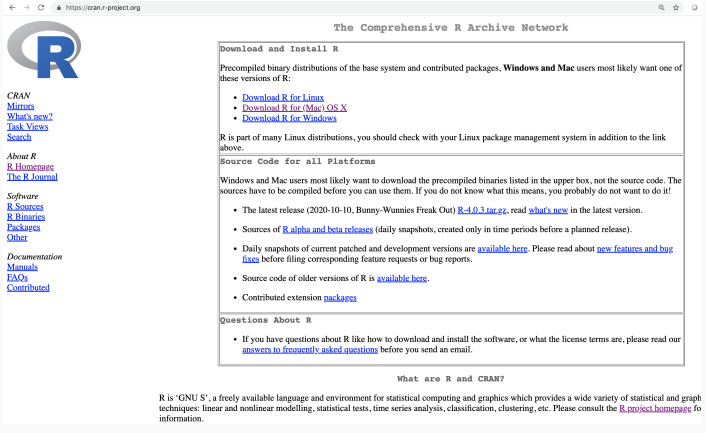
Views my own, not of employer

Introduction

- US Army (Automated Logistical Specialist) $\longrightarrow$ Database (Data Entry)
- UNC (BS in Mathematical Decision Science) $\longrightarrow$ Matlab
- New York Life Insurance (Actuarial Intern) $\longrightarrow$ Excel (Shortcut), Database (Query)
- Georgia Tech (MS/PhD in Industrial Engineering) $\longrightarrow$ Matlab
- EPA (Physical Scientist Intern) $\longrightarrow$ Database (MS Access)
- UPS (Security Analyst in Corp. Security) $\longrightarrow$ R (Plotting), Database (Data Warehouse)
- AT&T (Data Scientist in Chief Data Office) $\longrightarrow$ R (Packaging), Python, Big Data

Motivation

- CRAN
- install.packages("ggplot2")
- R package development workshop in 2017
- uncmbb package on CRAN
- The Carpentries
- Things I wish I had learned in school
- Some didn't exist, but mostly I just didn't know better
- Introductory by design, not comprehensive



Add Survey Result

Operating System

- Mainly for Windows users
- Local Machine (e.g., your computer) vs. Remote Server (e.g., school computing server)
- Know there are other operating systems
- Play with other operating systems (mainly Linux)
- There are many flavors of Linux, but don't be discouraged! (Ubuntu is just fine)
- Windows Subsystem for Linux (WSL)
- Try (Virtual Box, USB boot)

Shell

- Terminal
  - Really, a terminal *emulator*
  - A graphical window
  - Lets you interact with your operating system through shell
- Shell
  - Command line interface (CLI)
  - Scripting/programming language
  - Bash ("**B**ourne **a**gain **sh**ell") is default for many OS
- Terminal $\longrightarrow$ Shell $\longrightarrow$ Operating System
- Files, files, and more files
- Project directory structure
- Easier in action than in text

Text Files

- Most work in shell is text-based
- A variety of text editors
  - Vim
  - Emacs
  - Notepad/Notepad++
  - Visual Studio Code
  - Sublime
  - RStudio
  - And more
- Pick a text editor and try using it for any text-based tasks
  - Coding
  - Note taking
  - Presentation
- How to write in a text editor? `⟶` check out R Markdown

Languages of Data Science

- R or Python? Both!
  - "R is a language and environment for statistical computing and graphics"
  - "Python is a programming language that lets you work quickly and integrate systems more effectively"
- Plotting
  - Bar chart
  - Line chart
  - Covers majority of plotting needs
- Packaging
  - R Package
  - Python Package
  - Start w/ data package (babynames, uncmbb)
- And everything between plotting and packaging

Data Example

```
#install.packages("uncmbb") # if not already installed
library(uncmbb)
tail(unc)
```

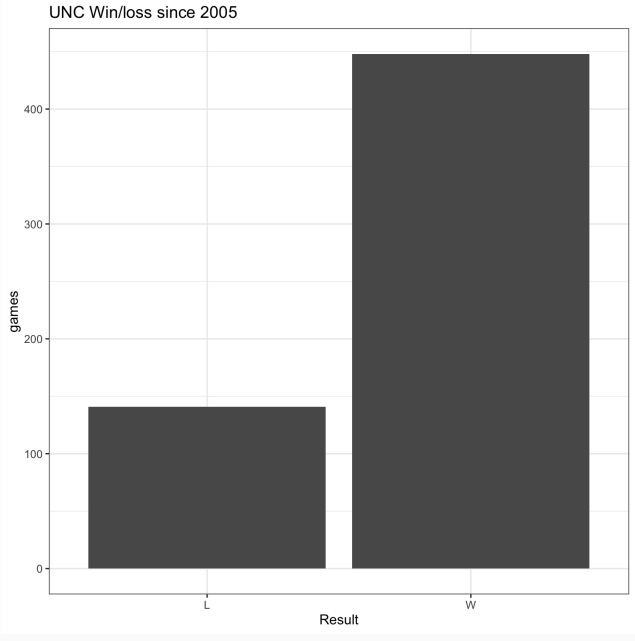
```
##      Season  Game_Date Game_Day   Type Where      Opponent_School Result Tm Opp   OT
## 2256   2020 2020-02-25      Tue    REG     H North Carolina State      W 85  79 <NA>
## 2257   2020 2020-02-29      Sat    REG     A             Syracuse      W 92  79 <NA>
## 2258   2020 2020-03-03      Tue    REG     H          Wake Forest      W 93  83 <NA>
## 2259   2020 2020-03-07      Sat    REG     A                 Duke      L 76  89 <NA>
## 2260   2020 2020-03-10      Tue CTOURN     N        Virginia Tech      W 78  56 <NA>
## 2261   2020 2020-03-11      Wed CTOURN     N             Syracuse      L 53  81 <NA>
```

```
tail(duke)
```

```
##      Season  Game_Date Game_Day Type Where      Opponent_School Result  Tm Opp   OT
## 2253   2020 2020-02-19      Wed  REG     A North Carolina State      L  66  88 <NA>
## 2254   2020 2020-02-22      Sat  REG     H        Virginia Tech      W  88  64 <NA>
## 2255   2020 2020-02-25      Tue  REG     A          Wake Forest      L 101 113  2OT
## 2256   2020 2020-02-29      Sat  REG     A             Virginia      L  50  52 <NA>
## 2257   2020 2020-03-02      Mon  REG     H North Carolina State      W  88  69 <NA>
## 2258   2020 2020-03-07      Sat  REG     H       North Carolina      W  89  76 <NA>
```

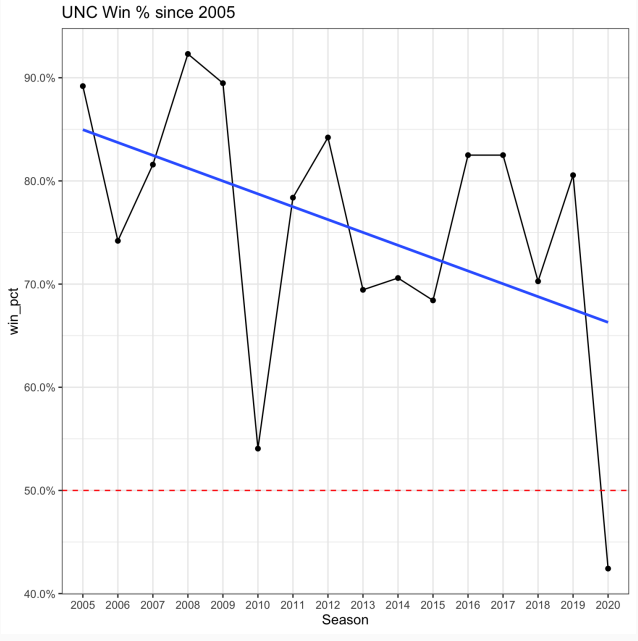
Bar Chart Example

```r
library(uncmbb)
library(dplyr)
library(ggplot2)

# prepare data for plotting
dat <- unc %>% filter(Season >= 2005) %>%
               group_by(Result) %>%
               summarize(games = n())

# plot aggregated data
dat %>% ggplot(aes(x = Result, y = games)) +
        geom_bar(stat = "identity") +
        labs(title = "UNC Win/loss since 2005")
```



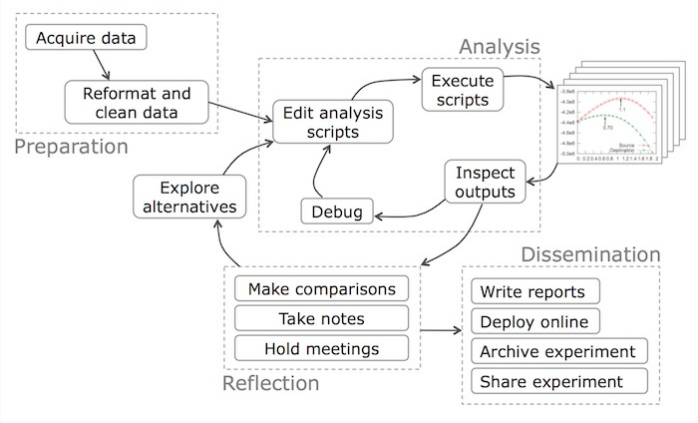
Line Chart Example

```r
library(uncmbb)
library(dplyr)
library(ggplot2)

# prepare data for plotting
dat ← unc %>% filter(Season ⩾ 2005) %>%
               group_by(Season) %>%
               summarize(games = n(),
                         wins = sum(Result == "W"),
                         losses = sum(Result == "L"),
                         win_pct = wins/games)

# plot aggregated data
dat %>% ggplot(aes(x = Season, y = win_pct, group = 1)) +
        geom_line() +
        geom_point() +
        geom_smooth(method = "lm", se = FALSE) +
        geom_hline(yintercept = 0.5,
                   linetype = "dashed", colour = "red") +
        scale_y_continuous(labels = scales::percent) +
        labs(title = "UNC Win % since 2005")
```



Data Science Workflow

- Example data science workflow (source)
- Missing, but important: **Problem Formulation**
- Iterative in nature
- Emphasis on "Analysis" step in school
- More emphasis on other steps in industry
- Team sport
  - Team lead
  - Project managers
  - Data engineers
  - Data scientists



Parting Thoughts

- In a nutshell, try
  - Ubuntu
  - Bash shell
  - Text editor
  - Bar/line charts in R/Python
  - Package things up in R/Python
  - Data science workflow
- Other topics that are not covered
  - Git (version control)
  - SQL
  - Blogging
  - Communication
  - Much more...

Links

- Good Enough Practices in Scientific Computing
- Carpentries Lesson on Shell
- Happy Git and GitHub for the useR
- Data Science at Command Line
- Editor War
- R for Data Science
- What They Forgot To Teach You About R
- R Graphics Cookbook
- Python Data Science Handbook
- Anaconda Data Science Toolkit
- Project-Oriented Workflow
- Why Jupyter Is Data Scientists' Computational Notebook of Choice
- The First Notebook War
- I Don't Like Notebooks

Questions?

Thank You!

In the future,

if any of the things in this talk ends up helping you in any way,

please reach out and let me know!

For now,

please let me know how the presentation was

by filling out the survey below!