**Introduction to Programming in R**

**A Term Fall 2021**

|  |  |
| --- | --- |
|  | [**Jared P. Lander**Links to an external site.](http://www.gsb.columbia.edu/cbs-directory/detail/jpl2135)  *Adjunct Assistant Professor* Decision, Risk, and Operations |

|  |  |
| --- | --- |
| **Email:** | [jpl2135@columbia.edu](mailto:jpl2135@columbia.edu) |

Textbook[**COURSE READINGS**](https://courseworks2.columbia.edu/courses/97660/pages/course-readings)

[*R for Everyone*, Second Edition (Links to an external site.)](http://amzn.to/2upiUrA)

[*Programming Skills for Data Science* (Links to an external site.)](https://amzn.to/2T0gPLI) -- suggested reading

The reading schedule is posted in the [Pages Section](https://courseworks2.columbia.edu/courses/97660/pages/course-readings).

Homework

Homework will be given on a weekly basis and is due at the start of the following class. Each assignment will be worth successively more points. You may work together on the homework, but your submission must be your own work and unique to you.

There is no grading penalty for the first late assignment. Subsequent late assignments will receive a 5% deduction, then 10%, for a max of 15% deduction on the fourth and any additional late assignments. Late assignments must be submitted within one week of their original due date.

Computing Environment

One of the biggest obstacles to learning a programming language is setting up your computing environment. To ease this process we have made a [Git repo (Links to an external site.)](https://github.com/jaredlander/cbs2020spring" \t "_blank). Please read the landing page carefully and follow the instructions before class.

R Syllabus

This class is an intensive introduction to R. It starts with the very basics of assigning variables and reading data. It then progresses to using RMarkdown for document and presentation creation. Following that it covers data manipulation, modeling, machine learning and dashboard creation.

Week 1

Introduction to R

* The RStudio Interface
* Basic Math
* Assigning Variables
* Working Directories
* Relative Paths
* Reading Data
  + Read from text files with [readr (Links to an external site.)](https://www.rdocumentation.org/packages/readr/" \t "_blank)
  + Read from Excel files with [readxl (Links to an external site.)](https://www.rdocumentation.org/packages/readxl/" \t "_blank)
* Writing Functions

RMarkdown

* [RMarkdown (Links to an external site.)](https://rmarkdown.rstudio.com/) Primer
  + Sections
  + Text Formatting
  + Lists
  + Links
* Integrating R into Markdown
  + Code Chunks
  + Chunk Options
* Including Figures
* Output Formats
  + HTML
  + PDF
  + Word
* Presentations

Week 2

Data Manipulation with [dplyr (Links to an external site.)](https://www.rdocumentation.org/packages/dplyr/" \t "_blank)

* Understanding a [tibble (Links to an external site.)](https://www.rdocumentation.org/packages/tibble/topics/tbl" \t "_blank)
* Use [pipes (Links to an external site.)](https://magrittr.tidyverse.org/) for cleaner code
* Select columns with [select (Links to an external site.)](https://www.rdocumentation.org/packages/dplyr/topics/select)
* Filter rows with [filter (Links to an external site.)](https://www.rdocumentation.org/packages/dplyr/topics/filter)
* Change and create columns with [mutate (Links to an external site.)](https://www.rdocumentation.org/packages/dplyr/topics/mutate)
* Calculate summary statistics with [summarize (Links to an external site.)](https://www.rdocumentation.org/packages/dplyr/topics/summarize)
* Group data for calculations with [group\_by (Links to an external site.)](https://www.rdocumentation.org/packages/dplyr/topics/group_by" \t "_blank)
* Joins with [left\_join (Links to an external site.)](https://www.rdocumentation.org/packages/dplyr/topics/left_join" \t "_blank)

Creating Visualizations

* [ggplot2 (Links to an external site.)](https://www.rdocumentation.org/packages/ggplot2/) paradigm
* Aesthetics
* Scatter plots
  + Color Coding
  + Size
  + Shape
  + Opacity
* Small multiple plots
* Histograms
* Density Plots
* Combining Layers
* Violin Plots
* Themes

Week 3

Reading Data

* CSVs with [readr (Links to an external site.)](https://www.rdocumentation.org/packages/readr/" \t "_blank)
* Databases with [DBI (Links to an external site.)](https://www.rdocumentation.org/packages/DBI/)
* JSON with [jsonlite (Links to an external site.)](https://www.rdocumentation.org/packages/jsonlite/" \t "_blank)
* Scraping web pages with [rvest (Links to an external site.)](https://www.rdocumentation.org/packages/rvest/" \t "_blank)

Iterate Over Lists with [purrr (Links to an external site.)](https://www.rdocumentation.org/packages/purrr/" \t "_blank)

* Basics of functional programming
* Mapping over a list
* Difference from [lapply (Links to an external site.)](https://www.rdocumentation.org/packages/base/topics/lapply" \t "_blank)
* Consistent Data Types
* Mapping to different data types
  + chacracter
  + numeric
  + data.frame
* Mapping functions with multiple arguments

Reshaping Data

* Convert from wide to long with [gather (Links to an external site.)](https://www.rdocumentation.org/packages/tidyr/topics/gather)
* Convert from long to wide with [spread (Links to an external site.)](https://www.rdocumentation.org/packages/tidyr/topics/spread)

Week 4

Linear Models

* Simple Linear Model with [lm (Links to an external site.)](https://www.rdocumentation.org/packages/stats/topics/lm)
* The Formula Interface
* Multiple Regression
* Tidying models with [broom (Links to an external site.)](https://www.rdocumentation.org/packages/broom/)
* Visualizing models with [coefplot (Links to an external site.)](https://www.rdocumentation.org/packages/coefplot/" \t "_blank)

Generalized Linear Models

* Logistic Regression for Binary Data
* Poisson Regression for Count Data
* Quasipoisson Regression for Overdispersed Count Data

Assessing Model Quality

* AIC
* BIC

Week 5

Cross-Validation

* Use Cross-Validation for Model Assessment

Penalized Regression

* L1 Penalty (Lasso)
* L2 Penalty (Ridge)
* Implement via the Elastic Net with [glmnet (Links to an external site.)](https://www.rdocumentation.org/packages/glmnet/" \t "_blank)
* Tuning Hyperparameters

Boosted Trees

* Decision Trees
* Boosted Trees
* Fit Model with [xgboost (Links to an external site.)](https://www.rdocumentation.org/packages/xgboost/" \t "_blank)

Week 6

[(Links to an external site.)](https://shiny.rstudio.com/)Forecasting Time Series Data

* Time Series formats
  + ts
  + xts
  + tsibble
* Visualizing Time series
* Benchmark Models
  + mean
  + naive
  + random walk
* Advanced Models
  + ETS
  + ARIMA
* Model Fitting
* Model Evaluation
  + AICc
  + Time Series Cross Validation
* Forecasting

Instructor Bio

Jared P. Lander is the Chief Data Scientist of [Lander Analytics](https://courseworks2.columbia.edu/courses/97660/assignments/www.landeranalytics.com), a data science and artificial intelligence consulting and training firm based in New York City; the organizer of the [New York Open Statistical Programming Meetup](https://courseworks2.columbia.edu/courses/97660/assignments/www.nyhackr.org)—the world’s largest R meetup—–and the [New York R Conference](https://courseworks2.columbia.edu/courses/97660/assignments/www.rstats.nyc)); author of [*R for Everyone*](https://courseworks2.columbia.edu/courses/97660/assignments/www.jaredlander.com/r-for-everyone) and an adjunct professor at [Columbia University](https://courseworks2.columbia.edu/courses/97660/assignments/www.columbia.edu). With an M.A. from [Columbia University](https://courseworks2.columbia.edu/courses/97660/assignments/www.columbia.edu) in statistics and a B.S. from [Muhlenberg College](https://courseworks2.columbia.edu/courses/97660/assignments/www.muhlenberg.edu) in mathematics, he has experience in both academic research and industry. Very active in the data community, Jared is a frequent speaker at conferences, universities and meetups around the world. His writings on statistics can be found at [jaredlander.com](https://courseworks2.columbia.edu/courses/97660/assignments/jaredlander.com) and his work has been featured in publications such as [Forbes (Links to an external site.)](https://www.forbes.com/sites/prishe/2017/03/07/reflections-from-the-2017-mit-sports-analytics-conference/#1a95a3473f75) and the [Wall Street Journal (Links to an external site.)](https://www.wsj.com/articles/a-data-scientist-dissects-the-2016-nfl-draft-1461793878).