

# Starting with R and data

Introduction to Data Science

Nina Zumel

John Mount

# Lesson Goals

- Learn how to start and work with the R analytics platform
- Load data and try simple calculations

**Required software**

# We recommend

- R: <http://cran.r-project.org>
- RStudio: <http://www.rstudio.com>

# This is not an R course

- This course requires a familiarity with R
- You *can* acquire such a familiarity by working through an R course and/or book in parallel with this course

# Data science in R is only a *small* subset of data science

- We are mostly teaching in an R context so we have a specific simple shared platform
- Most data scientists work using multiple platforms
- Other platforms include:
  - SAS
  - Python (pandas, scikit-learn)
  - Hadoop (Mahout)
  - SQL analytics
  - Microsoft Azure
  - And *many* others

# Starting with R and RStudio





# Try the help command

- Start R or RStudio and type `help(ls)` to get documentation on the `ls` command used in our example.

# The example data

From: [http://www.amstat.org/publications/jse/jse\\_data\\_archive.htm](http://www.amstat.org/publications/jse/jse_data_archive.htm)

Home prices set:

NAME: Modeling home prices using realtor data

TYPE: Random sample

SIZE: 76 observations, 19 variables

The article associated with this dataset appears in the Journal of Statistics Education, Volume 16, Number 2 (July 2008).

description: <http://www.amstat.org/publications/jse/datasets/homes76.txt>

set: <http://www.amstat.org/publications/jse/datasets/homes76.dat.txt>

Journal article: <http://www.amstat.org/publications/jse/v16n2/datasets.pardoe.html>

SUBMITTED BY:

Iain Pardoe

Lundquist College of Business

University of Oregon

1208 University of Oregon

Eugene, OR 97403

# Example Code

- Example code (and some of the slides) for this course is available from:

<http://winvector.github.io/IntroductionToDataScience/>

- Each lesson will remind you of the appropriate link as a resource.

# Additional resources

- R: <http://www.statmethods.net>
- RMarkdown: <http://rmarkdown.rstudio.com>
- Packages: <http://cran.r-project.org/web/views/>
- Books!

# Some books

- R programming
  - Norman Matloff *The Art of R Programming*
  - Garrett Golemund *Hands-On Programming with R*
- R plus statistics
  - Robert Kabacoff *R in Action*, 2nd edition
  - Jared P. Lander *R for Everyone*
- Data Science
  - Cathy O'Neil, Rachel Schutt *Doing Data Science*
  - Nina Zumel, John Mount *Practical Data Science with R*
- Machine Learning
  - James et. al. *An Introduction to Statistical Learning*
  - Hastie et. al. *The Elements of Statistical Learning*

# What you should now know

- Where to get R and RStudio
- How to perform basic operations in R
- How to load data into R
- Where to find more resources