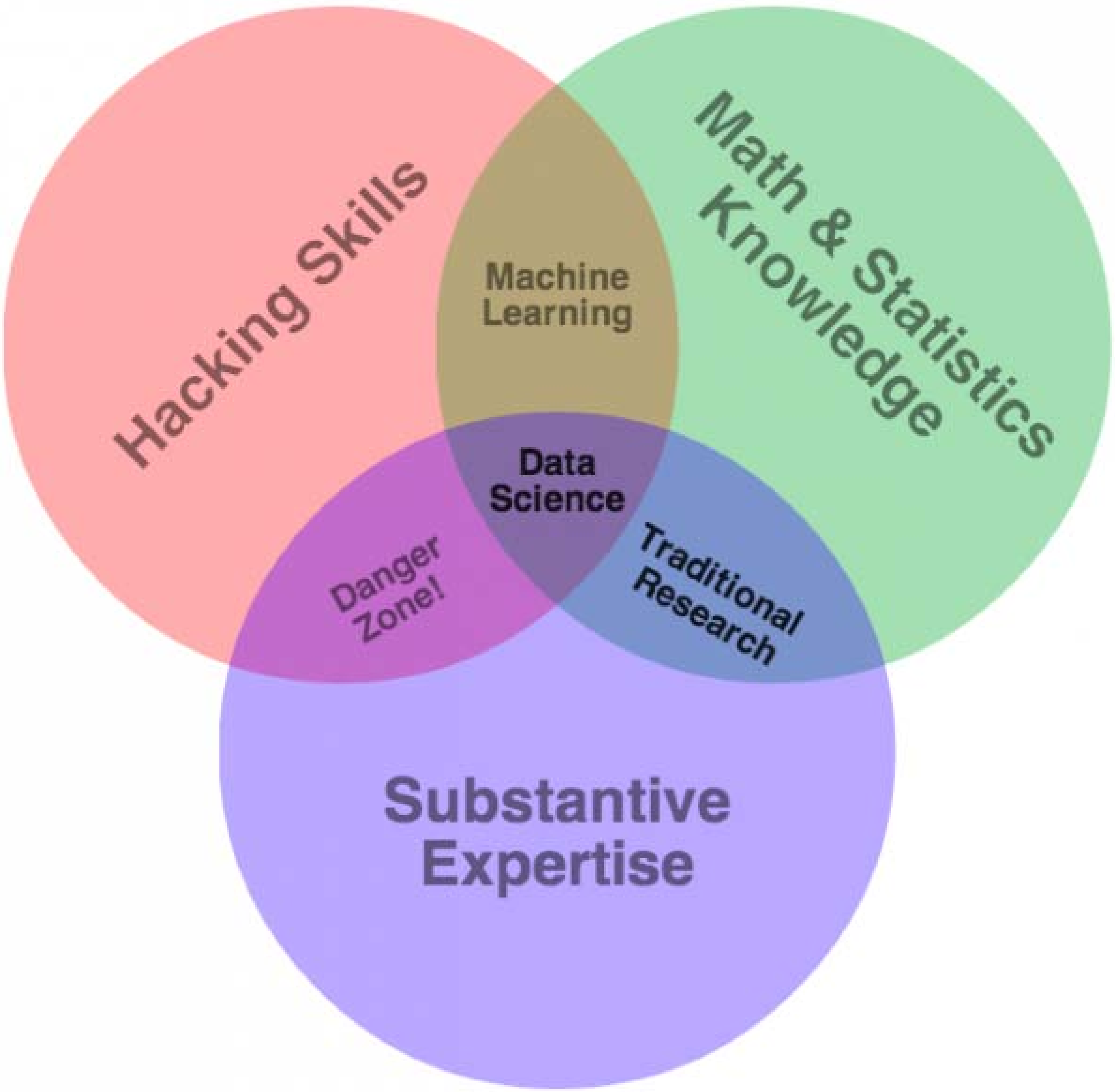
UNIVERSITY *of* WASHINGTON

Data Science UW
Methods for Data Analysis

Introduction and Data Exploration
Lecture 1
Stephen Elston


Hacking Skills
Math & Statistics Knowledge
Machine Learning
Data Science
Danger Zone!
Traditional Research
Substantive Expertise

Course Purpose

- This course focuses on essential concepts
- We are building foundations for your data science skills
- Course Objectives:
  - Learn methods to explore and understand data.
  - Understand the core concepts of probability and statistics.
  - Describe and interpret analytical results from common statistical methods.
  - Understand the mathematical basis of machine learning
  - Expand R programming skills to be able to write/test/log code from scratch.
  - Work with structured and unstructured data.
- See syllabus for more information:
  - https://canvas.uw.edu/courses/1105274/pages/datasci-350-b-course-syllabus

Course Requirements and Grading

This course will be graded by attendance, homework, and an individual project.

- Attendance: You MUST attend at least 6 out of 10 classes. **This is a non-negotiable UW requirement**.
- Homework must be completed by the start of the next class. (Assigned weeks 1-8).
  - Returned as a 0,1, or 2.
    - 0 = Not done or a major parts missing.
    - 1 = Completed, but missing or serious errors.
    - 2 = Completed with at most minor issues. Demonstrates full understanding of subject.
- Individual Project: Due at the start of the last class.
  - Counts as 8 points.

Course Requirements and Grading

There is a total of 24 possible points. (16 pts for hmk + 8 project)

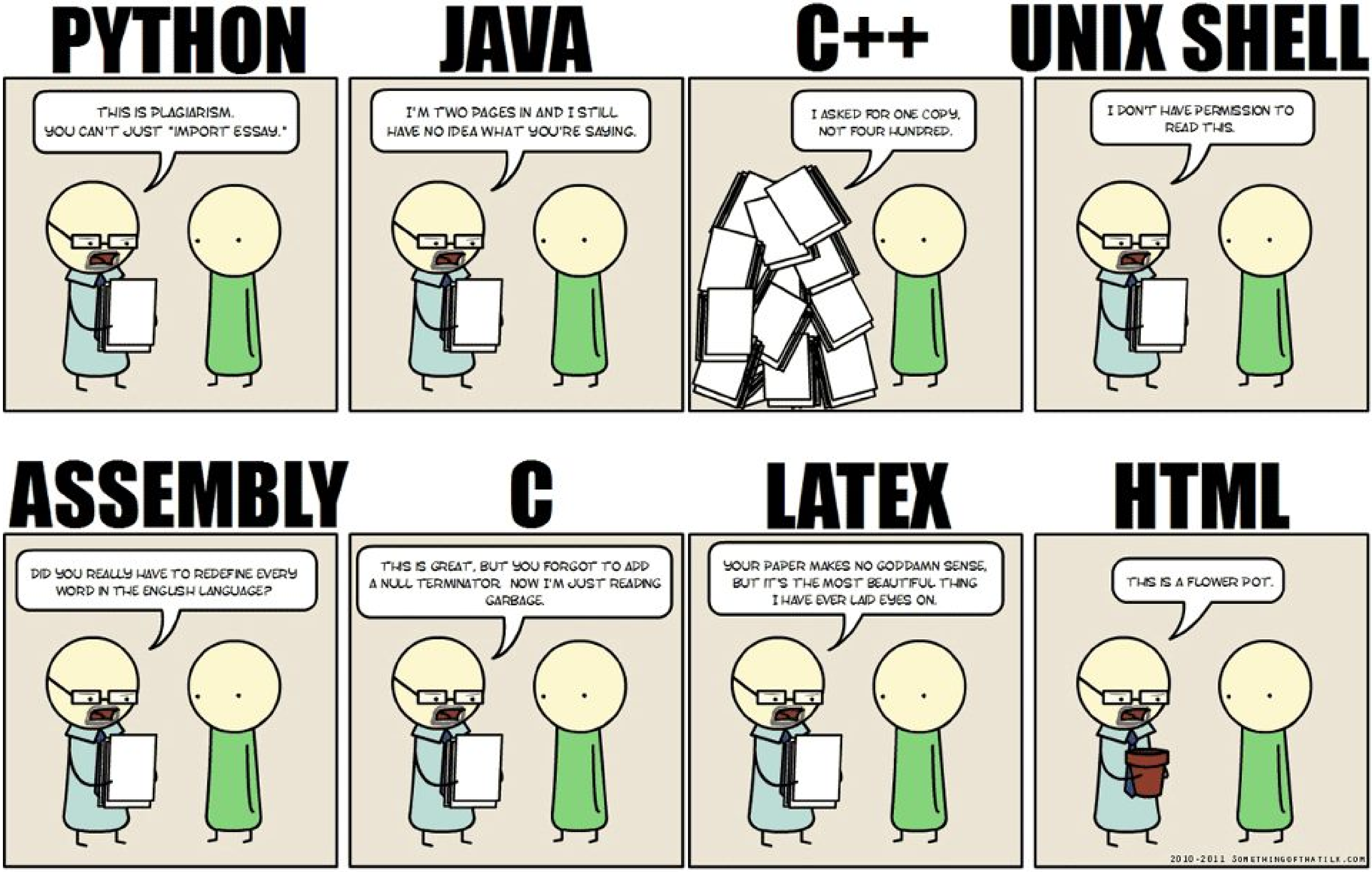
- Must get 18 total points to pass.
- All homework assignments must use good R coding technique
- Results must be presented in a professional style
- The individual project must be production level code.

Office Hours and Contact Information

- Contact me at:
  - stephen.elston@quantia.com
- When I'm *usually* available:
  - Off/on for simple things during work. (M-F 8am-5pm PST)
  - Sunday various afternoon/evening times.

Emergency contact: 402-980-3192

Review


PYTHON
THIS IS PLAGIARISM. YOU CAN'T JUST "IMPORT ESSAY."
JAVA
I'M TWO PAGES IN AND I STILL HAVE NO IDEA WHAT YOU'RE SAYING.
C++
I ASKED FOR ONE COPY, NOT FOUR HUNDRED.
UNIX SHELL
I DON'T HAVE PERMISSION TO READ THIS.
ASSEMBLY
DID YOU REALLY HAVE TO REDEFINE EVERY WORD IN THE ENGLISH LANGUAGE?
C
THIS IS GREAT, BUT YOU FORGOT TO ADD A NULL TERMINATOR. NOW I'M JUST READING GARBAGE.
LATEX
YOUR PAPER MAKES NO GODDAMN SENSE, BUT IT'S THE MOST BEAUTIFUL THING I HAVE EVER LAID EYES ON.
HTML
THIS IS A FLOWER POT.
2010-2011 SOMETHINGOFTHATILK.COM

Languages for data science

- Skills every data scientist should have
- SQL is the ‘Linqua Franka’ of data access
- R – widely used for visualization, statistical analysis, and machine learning
  - We use R in this course
- Python 3 – widely used for visualization, machine learning, and big data APIs (e.g. Spark)
  - Example for visualization: https://github.com/Quantia-Analytics/DyDataSF2016Visualization

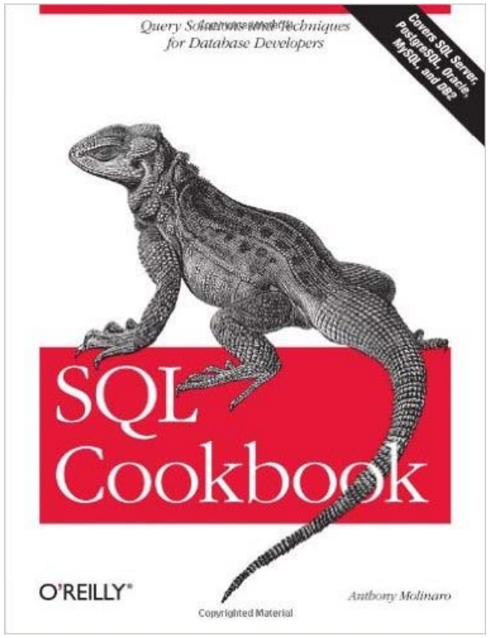
SQL Resources

SQL Tutorial and Resources

http://www.w3schools.com/sql/

Querying with Transact SQL Course, Graeme Malcom

https://www.edx.org/course/querying-transact-sql-microsoft-dat201x-3


Query Solutions and Techniques for Database Developers
Covers SQL Server, PostgreSQL, Oracle, MySQL, and DB2
SQL Cookbook
O'REILLY
Anthony Molinaro
Copyrighted Material

Prepare for R Demos

> Install R

https://cran.r-project.org/

-or-

https://mran.revolutionanalytics.com/download/

> Install RStudio

https://www.rstudio.com/products/rstudio/download/

GitHub

> Code, data and slides for this course are in a GitHub repository

https://github.com/StephenElston/DataScience350

> Install GitHub for desk top

https://help.github.com/desktop/guides/getting-started/installing-github-desktop/

- Or, just download the zip files

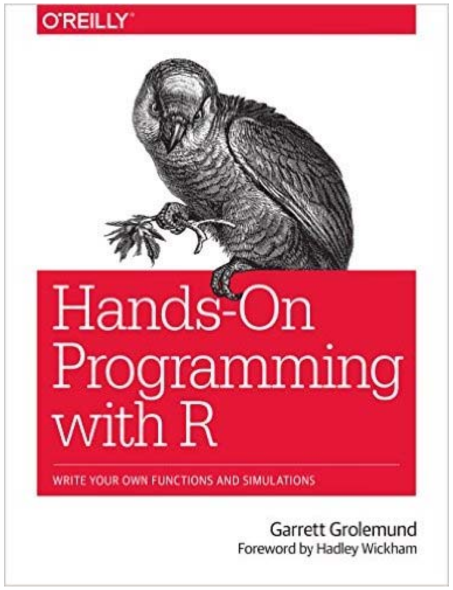
R Review

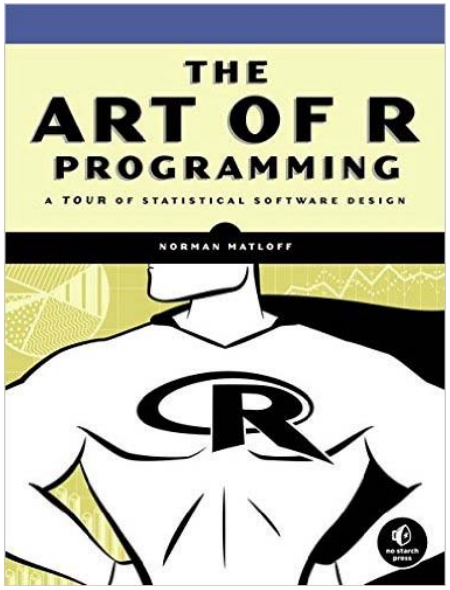
- R resources:
  - R page:
    - http://www.r-project.org/other-docs.html
  - Stackoverflow:
    - http://www.stackoverflow.com
  - ‘Little’ R intro:
    - http://cran.r-project.org/doc/contrib/Rossiter-RIntro-ITC.pdf
  - Quick R:
    - http://statmethods.net/
  - There are many tutorials available online, e.g.,
    - http://cyclismo.org/tutorial/R/
  - Google’s Style Guide:
    - http://google-styleguide.googlecode.com/svn/trunk/google-r
      style.html

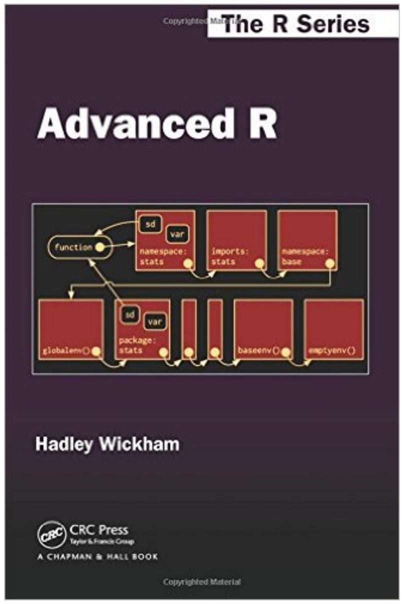
More R Resources

R Inferno, Pat Burns

http://www.burns-stat.com/pages/Tutor/R_inferno.pdf


O'REILLY
Hands-On Programming with R
WRITE YOUR OWN FUNCTIONS AND SIMULATIONS
Garrett Grolemund
Foreword by Hadley Wickham



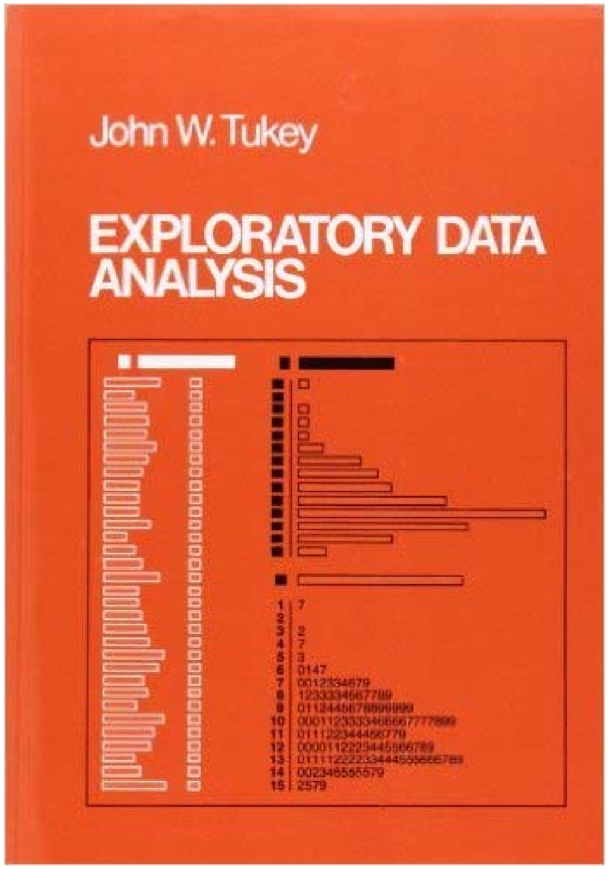


Exploratory data analysis

- Iterative exploration of the data with visualization
- Understand the relationships in the data
- Use multiple views of data
- Aesthetics to project multiple dimensions
- Conditioning to project multiple dimensions

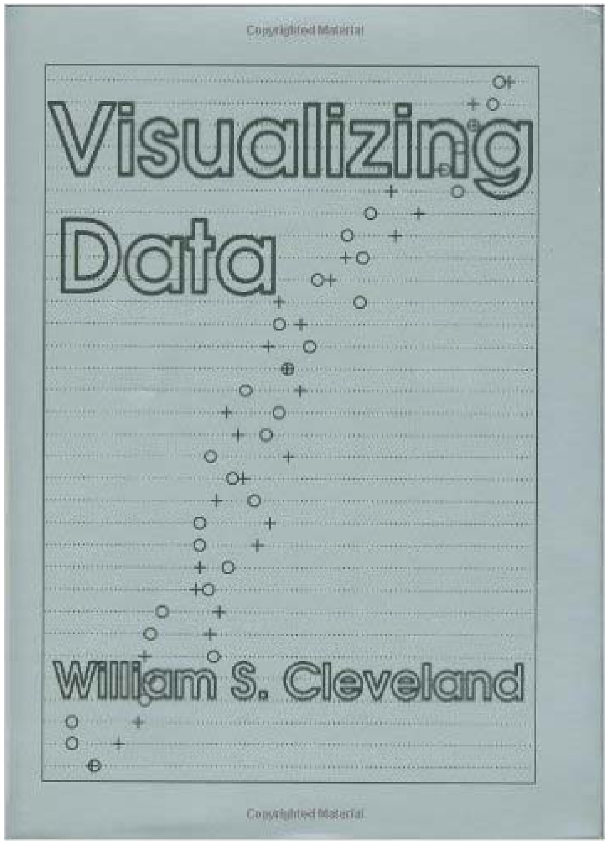
Seminal Book

John Tukey, Exploratory Data Analysis, 1977, Addison-Westley



Seminal Book

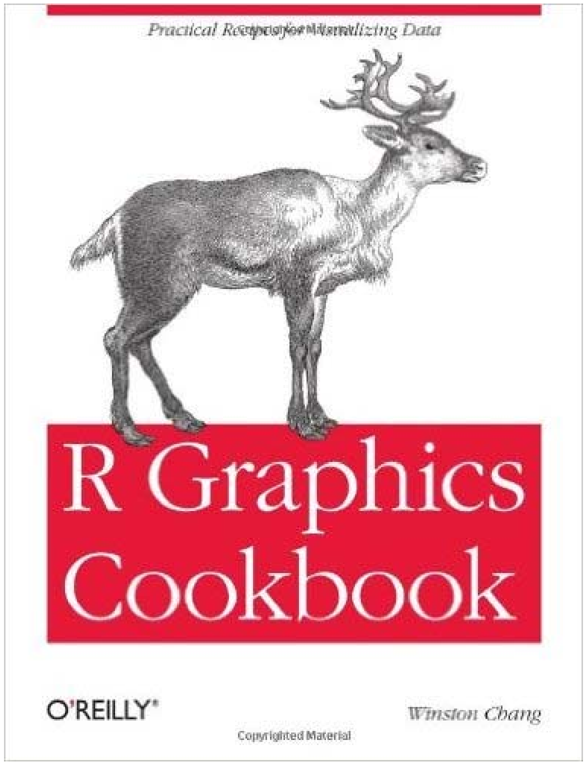
Visualizing Data, William S. Cleveland, Hobart Press 1993



ggplot2 resources

ggplot2 cheat sheet

https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf



R Demo

Data Visualization