Introduction to Data Science

Daniel Gutierrez, Data Scientist

Los Angeles, Calif.

Course Outcomes

- The ability to perform numeric exploratory data analysis (EDA)

Lesson Objectives

- Use numeric Exploratory Data Analysis (EDA) for knowledge discovery and statistical analysis
- Perform simple data analysis
- Use basic R statistical functions
- Explore levels of factor variables (categorical)
- Find number of non-missing values
- Independent study: common statistical tests for continuous random variables, and discrete data (categorical)

Simple Data Analysis

- Calculate unique values found for a variable (counts) using R functions and also SQL
- Summary statistics using `summary()` and `str()`
- Examining a data sample using `head()` and `tail()`

R Statistical Functions

- Calculate mean, min, max, range using `mean()`, `min()`, `max()` and `range()` respectively
- Calculate quantiles using `quantile()` and `fivenum()`
- Calculate variance using `var()`
- Calculate correlation using `cor()`
- Viewing a simple data distribution with `stem()`
- Calculate a cumulative sum with `cumsum()`

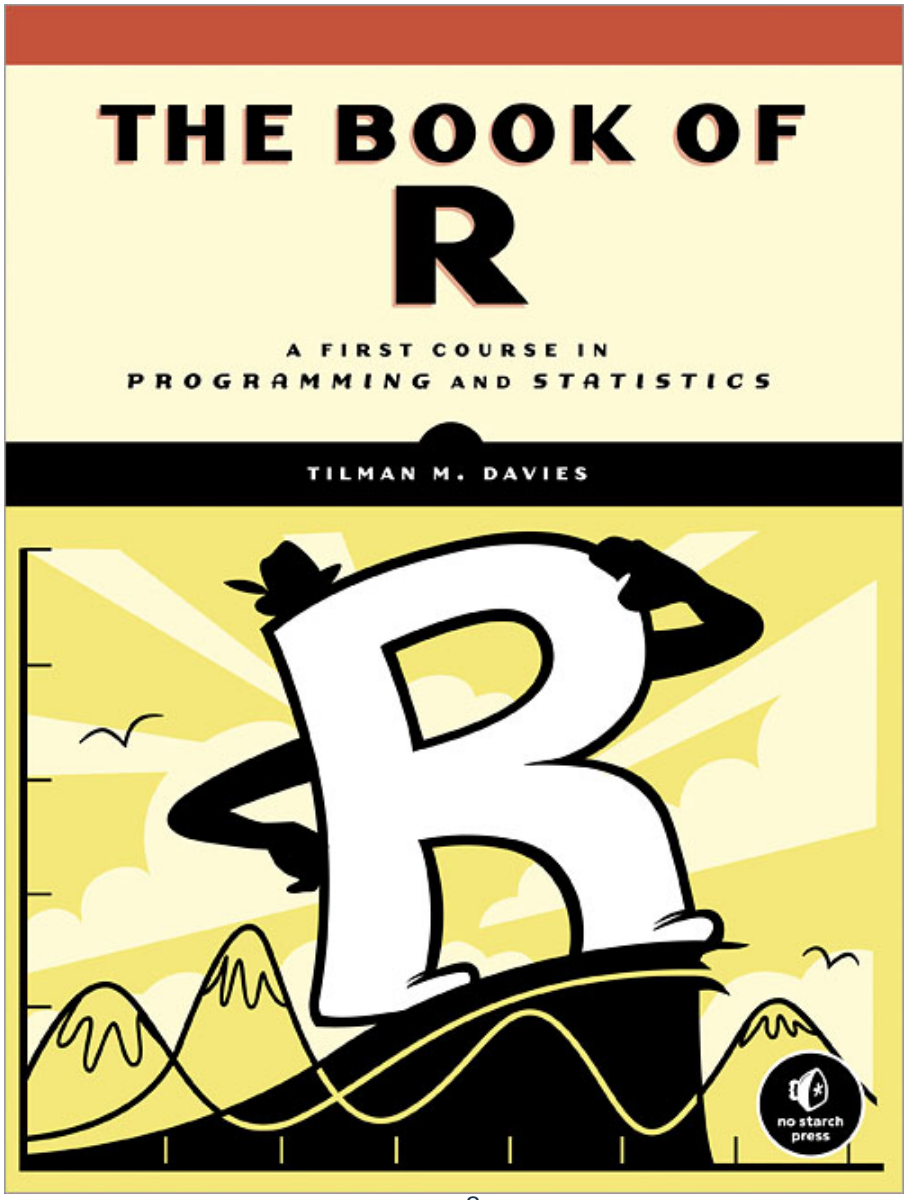
Exploring factor variables and NAs

- Explore levels of a factor variable (categorical) using `levels()`
- Produce a contingency table to count instances for each level in a factor variable. Can use `table()`
- Count non-missing values of a variable

Independent Study: Common Statistical Tests

- Common statistical tests for continuous random variables
- Common statistical tests for discrete data (categorical)

Independent Study: Common Statistical Tests



Independent Study: Common Statistical Tests

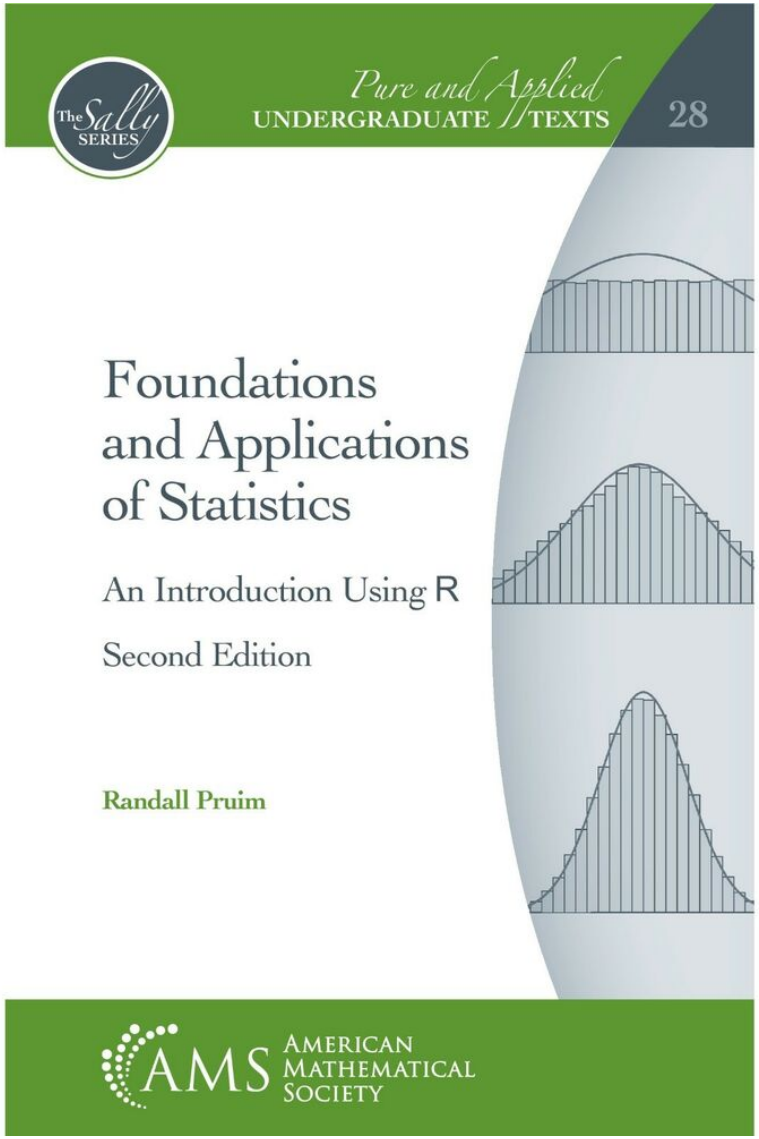
PART III: STATISTICS AND PROBABILITY

Chapter 13: Elementary Statistics
Chapter 14: Basic Data Visualization
Chapter 15: Probability
Chapter 16: Common Probability Distributions

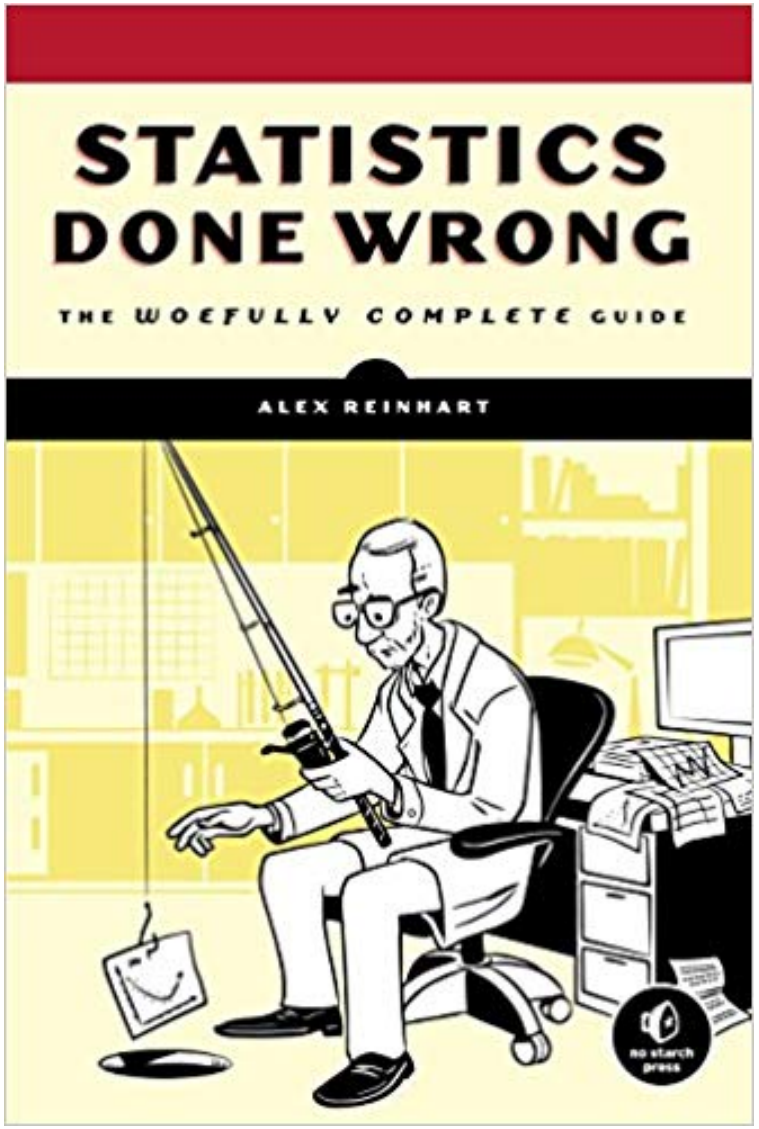
PART IV: STATISTICAL TESTING AND MODELING

Chapter 17: Sampling Distributions and Confidence
Chapter 18: Hypothesis Testing
Chapter 19: Analysis of Variance
Chapter 20: Simple Linear Regression
Chapter 21: Multiple Linear Regression
Chapter 22: Linear Model Selection and Diagnostics

Independent Study: Common Statistical Tests



Independent Study: Common Statistical Tests



Independent Study: Common Statistical Tests

Code module

- WEEK 7-1 Code module – Simple data analysis
- WEEK 7-2 Code module – R statistical functions, exploring factor variables and NAs
- WEEK 7-3 Code module – Independent study: common statistical tests

Summary

- In WEEK 7 of Introduction to Data Science, we built up our toolbox of EDA methods in order to gain familiarity with a data set.
- The methods discussed represent a small sample of available techniques. As you progress as a data scientist, you'll pick up more statistics that will help out in this step of the data science process.