Exploratory Data Analysis Notes

Coursera Course by John Hopkins University

INSTRUCTORS: Dr. Jeff Leek, Dr. Roger D. Peng, Dr. Brian Caffo

Contents

Intro

Instructor's Note

This course covers the essential exploratory techniques for summarizing data. These techniques are typically applied before formal modeling commences and can help inform the development of more complex statistical models. Exploratory techniques are also important for eliminating or sharpening potential hypotheses about the world that can be addressed by the data. We will cover in detail the plotting systems in R as well as some of the basic principles of constructing data graphics.
We will also cover some of the common multivariate statistical techniques used to visualize high-dimensional data.

All the best,

Roger Peng

Introduction

- EDA allows you to develop a rough idea of what your data look like and what kinds of questions might be answered by them.
- EDA is often the "fun part" of data analysis, where you get to play around with the data and explore.

- These techniques for summarizing data are typically applied before formal modeling commences and can help inform the development of more complex statistical models.

Exploratory Data Analysis with R Book

- **Exploratory Data Analysis with R**

The Art of Data Science

- **The Art of Data Science eBook**

- **The Art of Data Science printed version**

Installing R on...

Windows

- Just go to **the cran site** and install the Windows version.
  + For an optimal experience, back up all of data onto a usb, then install your prefered version of Linux (I use Fedora) and install the Linux version instead.

Mac

- Just go to **the cran site** and install the Mac version.
  + If you don't have enough money to buy a Mac install Linux instead, it's open-source, meaning it's free!

Installing R Studio (Mac)

- Install from **the RStudio website** *after* you have `R` installed.

Setting Your Working Directory on...

Windows, Mac & Linux

- Your working directory is where R will look for all the files it reads and where all the files it writes will go
- `getwd()` will display your current working directory
- `dir()` will display all files in your wd
- `setwd(param)` will set your working directory to the character string that is represented by `param`
- `source("myFunction.R")` will load in `myFunction` script from wd and any functions that are within it.

Lesson 1: Graphs

Principles of Analytic Graphics

Lesson with `swirl()`: Principles of Analytic Graphs

Exploratory Graphs

Lesson with `swirl()`: Exploratory Graphs

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 2: Plotting

Plotting Systems in R

Base Plotting System

Base Plotting Demonstration

Lesson with `swirl()`: Plotting Systems

Lesson with `swirl()`: Base Plotting System

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 3: Graphics Devices

Graphics Devices in R

Lesson with `swirl()`: Graphics Devices in R

Quiz 1 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Course Project 1

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 4: Lattice Plotting

Lattice Plotting System (part 1)

Lattice Plotting System (part 2)

Lesson with `swirl()`: Lattice Plotting System

Lesson with `swirl()`: Working with Colors

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 5: ggplot2 <3

Part 1

Part 2

Lesson with `swirl()`: GGPlot2 Part 1

Part 3

Part 4

Lesson with `swirl()`: GGPlot2 Part 2

Part 5

Lesson with `swirl()`: GGPlot2 Extras

Quiz 2 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 6: Hierarchical Clustering

Part 1

Part 2

Part 3

Lesson with `swirl()`: Hierarchical Clustering

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 7: K-Means Clustering & Dimension Reduction

K-Means Clustering (Part 1)

K-Means Clustering (Part 2)

Lesson with `swirl()`: K Means Clustering

Dimension Reduction (Part 1)

Dimension Reduction (Part 2)

Dimension Reduction (Part 3)

Lesson with `swirl()`: Dimension Reduction

Lesson with `swirl()`: Clustering Example

Reminder to commit to GitHub (Delete this line AFTER the commit)

Lesson 8: Working with Color in R Plots

Part 1

Part 2

Part 3

Part 4

Quiz 3 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Case Studies

Clustering Case Study

Air Pollution Case Study

Lesson with `swirl()`: CaseStudy

Quiz 4 Scribbles

Reminder to commit to GitHub (Delete this line AFTER the commit)

Course Project 2

Reminder to commit to GitHub (Delete this line BEFORE the commit)